Universidad de Valladolid

Departamento de Informática

TESIS DOCTORAL

Recuperación de Información en Persa: Revisión Crítica y Propuestas de Mejora

Presentada por Mohammad Sadeghi Hassanabadi
para optar al grado de doctor por la Universidad de Valladolid

Dirigida por: Dr. D. Jesús M. Vegas Hernández

Valladolid, Mayo de 2015

Resumen

Este trabajo de tesis se enmarca en la disciplina de la Recuperación de Información, RI cuyo objetivo es el de identificar, en una colección de documentos, aquellos que son relevantes a una necesidad de información del usuario.

Los principales modelos y técnicas propuestas por los investigadores en la recuperación de información dependen en mayor o menor medida del idioma, tanto de los documentos como de las consultas que se formulan. Al respecto existen muchos estudios, desde el procesamiento del texto hasta los modelos de recuperación, para las lenguas occidentales o *predominantes* como el inglés. Sin embargo, las lenguas minoritarias como la lengua persa no han sido suficientemente tratadas en los sistemas de RI tanto tradicional como Web y deberían ser mejoradas desarrollando algoritmos y técnicas que consideran las características de esas lenguas. Además, la diferencia morfológica y lingüística de persa frente al inglés en todos sus niveles no permite una extrapolación al idioma persa de los resultados obtenidos para el inglés, demandando la realización de trabajos específicos.

El objetivo principal de esta tesis es analizar la recuperación de información en persa y poner de manifiesto los factores que afectan a la eficiencia en los sistemas de recuperación de información con documentos persas y dar propuestas para mejorar la eficiencia de la recuperación de documentos relevantes.

Los resultados obtenidos por nuestros experimentos revelan que la mala representación de documentos es el gran desafío que presenta la lengua persa en un sistema de recuperación de información. La representación de documentos es el conjunto de operaciones que se hacen sobre el contenido de un documento desde su creación hasta su indexación.

Las diferentes formas de la escritura, ambigüedades en el texto escrito, la dispersión en posición alfabética diferente y no estándar de la ortografía son los principales problemas que hacen necesaria una pre-normalización o *estandarización* del texto para cumplir con los criterios de un sistema de RI.

La segmentación del texto y la definición del límite de las palabras son tareas muy difíciles en persa. Hay de uno a cuatro formas de escribir un carácter alfabético según su posición en una palabra. Cada forma puede ser inicial, media, final y aislada. Hay varias formas de escribir textos persas que difieren en el estilo de escritura de palabras usando o eliminando los espacios dentro o entre las palabras utilizando diversas formas de caracteres. Entonces la correcta tokenizatión y la conversión de estas formas y estilos en una única norma es un paso necesario en la construcción de los sistemas de RI con documentos persas.

En la lengua persa, las palabras se construyen generalmente a partir de la forma imperativa de los verbos. Por lo tanto, desde un punto de vista de la lingüística, la primera etapa para extraer la raíz es encontrar el modo imperativo de la palabra. En general, no es fácil obtener el modo imperativo ya que hay muchos infinitivos irregulares. La forma imperativa del infinitivo irregular se basa en cómo se escuchan o se usan las palabras. En este caso, se necesita buscar el modo imperativo en el léxico. Además, la diversidad de formas plurales, plurales irregulares y las palabras no plurales terminando con los signos plurales son algunos de los retos en la construcción de lematizadores para la lengua persa.

Otro objetivo de esta tesis es la evaluación del rendimiento y calidad de los buscadores de web frente a los documentos persas, en particular, de Google que es utilizado por el 92% de los usuarios iraníes. Los resultados obtenidos han destacado que el buscador Google considera las palabras vacías persas como palabras claves del contenido de un documento persa. Además, la tokenización del texto no realiza correctamente la separación adecuada de las palabras y la lematización contiene muchos errores. En conclusión, Google debe mejorar las operaciones que corresponden a la representación de documentos persas teniendo en cuenta de la estructura y gramática de la lengua persa.

Analizando los trabajos previos, hemos constatado que no hay ninguna investigación que consiga en identificar automáticamente las palabras vacías en un sistema de RI. Por lo tanto, la última aportación de este trabajo es desarrollar un método automático que permita identificar las palabras vacías para sistemas de RI con documentos persas. Nuestro método está basado en los modelos estadísticos y en el

modelo de información. El modelo estadístico extrae las palabras vacías teniendo en cuenta la distribución de estas palabras en un corpus y en cada documento del corpus. El modelo de información mide el significado de una palabra en el texto mediante el uso de la teoría de la información.

Abstract

This PhD is part of the Information Retrieval, IR whose goal consists of identifying the documents in a collection that are relevant to a user information need.

The main models and techniques proposed by researchers in IR depend, to a greater or lesser extent, on the language of both the documents and the query. In this regard, there are many studies, from text processing to retrieval methods, for western or dominant languages like English. However, the minority languages such as Persian language have not been sufficiently treated in both the traditional IR and the Web and should be improved by developing algorithms and techniques that consider the characteristics of those languages. Moreover, the difference morphological and linguistic of Persian with respect to English does not allow a direct extrapolation to the Persian language of the results obtained for English, demanding, in this way, the implementation of specific experiments.

The aim of this PhD is to analyze the information retrieval in Persian language and highlight the factors that affect the efficiency of information retrieval systems with Persian documents and make proposals to improve the efficiency of the relevant retrieved documents.

The obtained results by our experiments show that the wrong representation of the documents is the great challenge posed by the Persian language in the information retrieval system. The representation of documents is the set of operations performed on the content of a document from its creation to its indexing. The different forms of writing, ambiguities in written text, the dispersion in the different alphabetical position and non-standard spelling are the main problems that a pre-normalization or standardization of the text is required to comply with to adapt in an IR system.

Text segmentation and word boundaries are very difficult tasks in Persian. There are one to four written forms for each character according to its place in a word. Each form may be initial, medial, final and isolated. There are several ways of writing Persian texts that differ in the style writing, using or elimination of spaces and

using various forms of characters. Therefore, tokenization and conversion of these forms and styles to a unique standard is a necessary step in building an IR system with Persian documents.

In the Persian language, words are usually built up from the imperative forms of the verbs. Hence, from a linguistic point of view, the first step in extracting the root is to find the imperative mood of the word. In general, obtaining the imperative mood is not easy since there are irregular infinitives. The imperative form of irregular infinitives are based on how the words re heard or used. In this case, it needs to find the imperative mood in the lexicon. Moreover, the diversity of the plural forms, irregular plurals and plural words not ending with the plural signs are some of the challenges in building stemmers for the Persian language.

Another goal of this thesis is the performance and quality evaluation of the web search engines in the Persian documents, in particular Google, which is being used by almost 92% of Iranian users. The obtained results showed that the Google search engine considers the Persian stop words as the content keywords of a Persian document. In addition, the text tokenization does not perform correctly to separate the words from text and the stemming process contains many errors. In conclusion, Google should improve the operations corresponding to the representation of Persian documents by taking into account the structure and grammar of the Persian language.

Analyzing the previous works, we found that there is no research to identify automatically the Persian stop words for an IR system. Therefore, the last contribution of this work is to develop an automatic method to construct the stop words for Persian IR systems. Our method is based on statistical methods and information model. The statistical model extracts the stop words by considering the distribution of these words in a corpus and each document of corpus. The information model measures the meaning of a word in the text by using the information theory.

Agradecimientos

Tengo que agradecer en primer lugar a Jesús, mi director de tesis. Sin su ayuda, comprensión y paciencia no hubiera podido llegar al final. Gracias por iniciarme en la investigación y aportarme ideas que han resultado fundamentales en este trabajo.

Pienso que su papel en dirigir mi tesis fue más que un director de tesis. Recuerdo todas las reuniones semanales que hemos tenido juntos dedicando su tiempo para mí y cada vez me recibió en su despacho con una sonrisa lo que me dio el coraje de seguir mi trabajo. ¡No todo el mundo tiene la suerte de contar con gente así!

Desearía agradecer también a los miembros del tribunal el haber puesto a mi disposición su valioso tiempo y sabiduría para juzgar mi trabajo.

Tengo que agradecer a la Universidad de Valladolid por haberme dado la oportunidad de participar en el Doctorado de Informática.

Me gustaría agradecer a los profesores con quien he cursado el período de docencia y a los miembros del Departamento de Informática especialmente a Rosa, secretaria del departamento, por su amable cooperación durante estos años.

Siempre recordaré con agrado los buenos momentos que he pasado en España durante mis estancias en la Universidad de Valladolid. Fue una gran oportunidad de conocer la cultura y la lengua española aunque mi nivel de idioma es aún muy bajo en comparación con la riqueza de la lengua española.

Por último, un cariñoso agradecimiento a las personas más cercanas, mis amigos y mi familia.

Índice General

Índice de Tablas

Índice de Figuras

Parte I

Introducción y la Lengua Persa

Capítulo 1

Introducción

Resumen

Este capítulo constituye una breve presentación del concepto de la recuperación de información y especialmente la parte de la representación de los documentos en un sistema de recuperación de información. Al mismo tiempo se exponen la motivación de realizar esta tesis y también los objetivos y resultados esperados por los experimentos desarrollados durante este trabajo. En la última parte, se describe brevemente la estructura del presente documento, resumiéndose el contenido de cada capítulo.

1.1. Recuperación de información

A partir de la expansión y consolidación de Internet, como medio principal de comunicación electrónica de datos, se ha puesto a disposición de casi toda la humanidad una importante cantidad de información. A los efectos de aprovechar todo este potencial de información, es necesario poseer accesos que permitan que la tarea de recuperación sea eficiente y efectiva. Para dar respuesta a las necesidades de información de los usuarios, existen los que podríamos denominar, de forma genérica, sistemas de procesamiento automático de la información (Vilares Ferro, 2005). Estos sistemas tienen como objetivo suministrar, de forma eficaz y eficiente, aquella información solicitada por los usuarios.

Convendría entonces que las operaciones de dichos sistemas tomen en cuenta las características de la lengua en la cual están escritos los documentos y formuladas las consultas. Esto plantea un desafío interesante; hay importantes volúmenes de información a los cuales se debe acceder a petición de los usuarios, por lo tanto, cómo se pueden relacionar las preguntas con sus correspondientes respuestas, las necesidades de información con la documentación informatizada y las consultas con los resultados.

Bien, en las ciencias de la computación existe un área denominada Recuperación de Información, RI (*Information Retrieval*, en inglés), que estudia y propone soluciones al escenario presentado, planteando modelos, algoritmos y heurísticas.

Se puede comprobar, en el caso de una lengua minoritaria como la persa, que no ha sido suficientemente tratada por los sistemas de RI, tanto tradicional como Web y deberían ser mejorados, desarrollando algoritmos y técnicas que consideren las características y particularidades del idioma persa. Las investigaciones llevadas a cabo en el campo de la RI se han centrado en las lenguas occidentales y mayoritariamente en el inglés, siendo muy escaso el trabajo realizado para mejorar la recuperación de información de documentos escritos en persa. De lo antes planteado surgió la hipótesis de que los documentos en persa suelen ser mal representados en los sistemas de RI y no cumplen con los criterios usualmente establecidos en dichos sistemas.

La representación de documentos o *vista lógica de documentos* es el conjunto de operaciones que se hacen desde la creación de un documento hasta la indexación de su contenido (Baeza-Yates, et al., 1999). Los términos de índice se pueden generar correctamente si esas operaciones consideran apropiadamente la morfología y las características de la lengua en que están escritos los documentos. La correcta indexación es uno de los métodos más recomendados para acceder a los datos contenidos en los documentos y se conoce como la solución más eficaz para recuperar elementos de información de manera rápida y precisa (Mayfield, et al., 2003).

En consecuencia, se propone como objetivo principal de este trabajo demostrar que la hipótesis planteada es cierta, es decir, la conversión de un documento en persa a un conjunto de términos de índice necesita tratamientos especiales y se deben hacer propuestas para mejorar la eficiencia en la recuperación de información relevante implementada por un sistema de RI con documentos en persa.

El objetivo de un sistema de RI es identificar aquellos documentos de la colección que son relevantes a una necesidad de información del usuario planteada mediante una consulta formada por una serie de términos o palabras claves (Baeza-Yates, et al., 1999). Como resultado, estos sistemas devuelven una lista de documentos

que suele presentarse ordenadamente en función de valores que tratan de reflejar en qué medida cada documento responde a dichas necesidades de información. El ejemplo más común de un sistema de este tipo son los motores de búsqueda en Internet. Para ilustrar el problema nos ayudamos de la Figura 1.1, adaptada de Rijsbergen (Van-Rijsbergen, 1979).

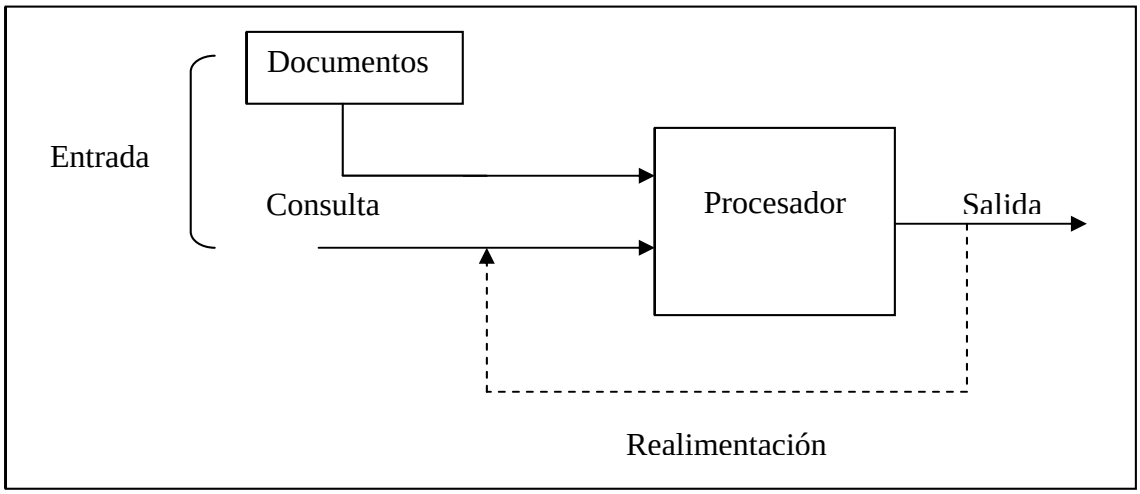

Figura 1.1: Sistema de recuperación de información

En esa figura se ha supuesto que el sistema de recuperación de información es una caja negra que acepta documentos y consultas, dando como resultado una salida, que es el conjunto de documentos que satisfacen la consulta. El problema principal en este sistema es obtener representaciones homogéneas de documentos y consultas, procesar convenientemente esas representaciones y obtener la salida. Por otro lado, también es muy importante la evaluación de la salida, para determinar si ésta coincide con las necesidades informativas del usuario.

En el proceso de recuperación de información se suelen distinguir las siguientes etapas:

- **Obtener representación de los documentos:** Generalmente los documentos se presentan utilizando un conjunto más o menos grande de términos de indexación. La elección de dichos términos es el proceso más complicado. Los tipos de cambios que se hacen sobre el contenido de los documentos en esta etapa son llamados “transformación de texto” o, más comúnmente,

"procesamiento de texto". El objetivo del tratamiento de textos es convertir las muchas formas en que las palabras pueden ocurrir en términos de índice más consistentes. Los términos de índice son la representación del contenido de un documento que se utilizan para la búsqueda (Croft, et al., 2009).

- **Identificar la necesidad de información del usuario:** Se trata de obtener la representación de esa necesidad, y plasmarla formalmente en una consulta acorde con el sistema de recuperación.
- **Búsqueda de documentos que satisfagan la consulta:** Consiste en comparar las representaciones de documentos y la representación de la necesidad informativa para seleccionar los documentos pertinentes.
- **Obtención de resultados y presentación al usuario.** Los documentos devueltos que corresponden a una consulta se muestran en una lista ordenada. El usuario puede hacer una selección de ellos considerando la relevancia de cada documento con respecto a su necesidad de información. La forma estándar de hacer esto es proporcionar un breve resumen del documento que está diseñado para permitir al usuario decidir sobre su relevancia (Manning, et al., 2008).
- **Evaluación de los resultados por parte del usuario.** Realizar los ajustes necesarios en el sistema basados en la realimentación con los usuarios para aumentar la calidad de la respuesta.

En un sistema de RI es muy importante la manera en que se almacena la información, así como el tipo gramatical (la importancia de información contenida) que tienen los términos dentro de los índices que describen a las colecciones. Esto puede ayudar a la eficiencia y precisión que pueden existir dentro del proceso de RI. A eso se debe que sea esencial entender un concepto básico, como la representación (o vista lógica) de los documentos.

Ya hemos mencionado anteriormente la representación de los documentos es una tarea que varía según el idioma y depende de la lengua en que están escritos los documentos. Aunque en este paso las operaciones son muy similares para todos los idiomas, las técnicas y algoritmos son diferentes y dependen, sobre todo, de la particularidad y propiedad del lenguaje de los documentos. La vista lógica se refiere a la manera en la que se representa un documento en un conjunto de índices. La forma

más sencilla de representar un documento es por medio del conjunto de palabras del texto completo, sin embargo este puede llegar a ser muy grande y por ello es conveniente reducirlo a una lista con las palabras clave del texto. De ahí se obtiene la primera forma de representación que se llama texto completo. Así las vistas lógicas pueden variar de acuerdo a los diferentes tipos de operaciones que se apliquen al texto.

Según Baeza-Yates (Baeza-Yates, et al., 1999), la vista lógica de un documento en un sistema de RI incluye algunas operaciones sobre su contenido que se puede ver en detalle en la Figura 1.2.

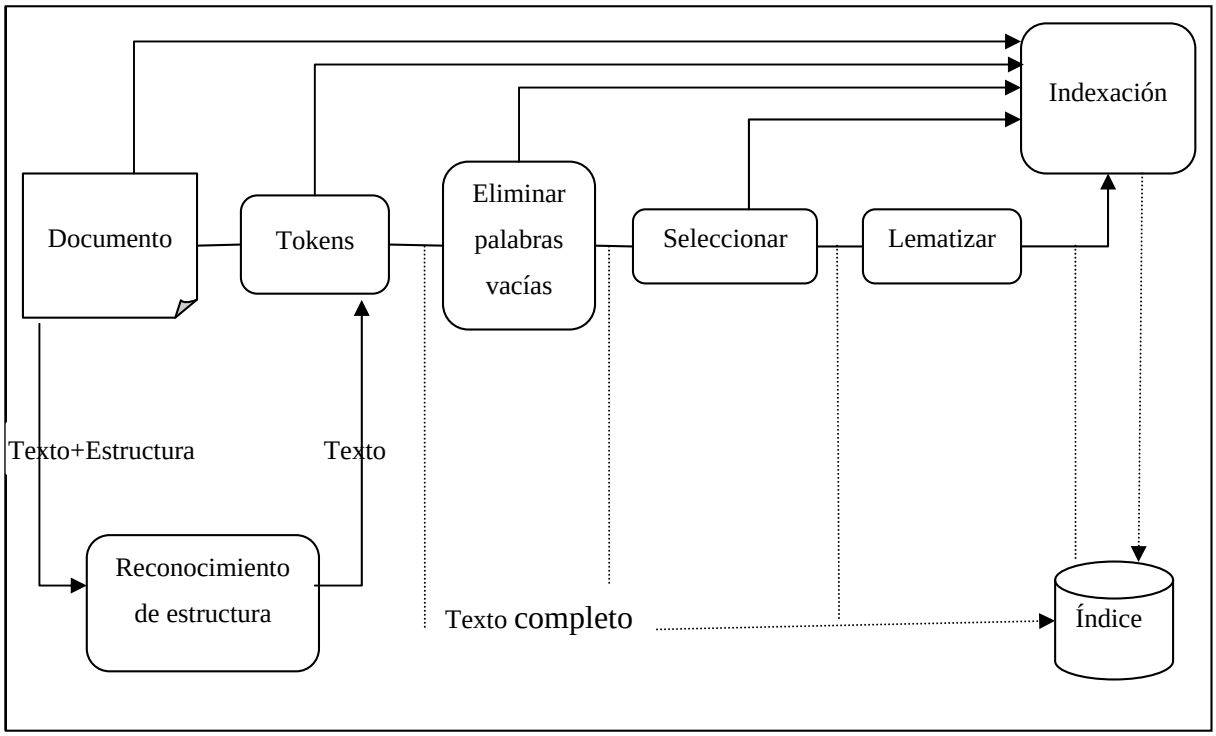

Figura 1.2: Procesamiento del texto para un sistema de RI, adaptada de (Baeza-Yates, et al., 1999)

En esta tarea, las operaciones más comunes en un sistema de RI son:

- **Tokenización**: Es el proceso de convertir un documento, visto como un flujo de caracteres, en palabras o elementos. Los elementos son grupos de caracteres con significado.

- **Eliminación de palabras vacías**: Consiste en eliminar las palabras que no dan una información importante sobre el contenido de documento.
- **Lematización**: Es el proceso que permite reducir el número de términos de indexación utilizando algún control morfológico o de las formas flexivas de las palabras.

Desde hace tiempo los ordenadores pueden representar un documento utilizando su contenido completo. Efectivamente, es la forma más completa de representar un documento, pero implica un coste computacional muy alto para colecciones grandes de documentos. En estos sistemas cada documento se puede representar por todas sus palabras, tanto nombres, como verbos, adjetivos, adverbios, etc. A pesar de ello, no todos los términos poseen la misma utilidad para describir el contenido de un documento. De hecho, hay términos más importantes que otros, pero no es tarea fácil decidir la importancia de cada término. Por ejemplo, desde el punto de vista de la recuperación de información existen palabras casi vacías de contenido semántico, como los artículos, preposiciones o conjunciones, que parecen poco útiles en el proceso.

Independientemente de ello, no son útiles en las tareas de recuperación de información aquellos términos que se repiten con mucha frecuencia en toda la colección de documentos, pues son términos poco discriminatorios en relación con una consulta dada. En conclusión, en estos sistemas no solo se persigue encontrar aquellos términos que mejor representen a los documentos, sino además aquellos que permitan diferenciar unos respecto de otros.

Por otra parte, hay sistemas que permiten realizar búsquedas conocidas como de texto libre. En estos se realizan búsqueda de sub-cadenas, sobre el texto almacenado de todo el documento. Evidentemente, el coste de almacenamiento y procesamiento computacional en la búsqueda es elevado, y además, muchas veces se muestran incapaces de resolver dos problemas básicos: la sinonimia y la polisemia. Por ello, la investigación en recuperación de información busca diseñar sistemas que acepten consultas en lenguaje natural y proporcionen documentos adecuados a tales consultas, ordenados según algún criterio del sistema, de acuerdo con las características de los

documentos y a la necesidad de información expresada por el usuario en su consulta (Belkin, et al., 1987).

En algunos casos, como sucede en la lengua persa, hay un paso adicional en el procesamiento del texto para un sistema de RI. Debido a sus ambigüedades en el texto escrito, la dispersión en posición alfabética diferente, la ambigüedad del límite de la palabra y la variedad de ortografía necesita experimentar una pre-normalización o preparación del texto que es una forma de estandarización de los términos. Esta operación que reduce la variedad de las palabras en favor de la uniformidad debe ser antes de proceder a la tokenización del texto.

1.2. Motivación

La recuperación de información es un área de investigación fuertemente relacionado con las ciencias de la documentación y con la informática. Los principales modelos y técnicas propuestas por los investigadores en RI dependen en mayor o menor medida del idioma, tanto de los documentos como de las consultas que se formulan. Al respecto existen muchos estudios, desde el procesamiento del texto hasta los modelos de recuperación, para las lenguas occidentales como el inglés, pero no se puede decir lo mismo en otros casos.

En este contexto, las lenguas minoritarias como la lengua persa no han conseguido atraer el interés de los investigadores a crear herramientas específicas para estos idiomas. Además, las diferencias morfológicas y lingüísticas del persa frente al inglés en todos sus niveles no permiten una extrapolación al idioma persa de los resultados obtenidos para el inglés, demandando la realización de trabajos específicos. Resulta paradójico, en este sentido, que siendo la lengua persa con una morfología completamente diferente del inglés, la investigación sobre RI en persa resulte tan exigua. Esto es cierto, desde luego, para lo que se refiere a la formulación de modelos generales teóricos, pero también para lo que es investigación experimental, en el sentido de la producción y ajuste de técnicas que implementen lo propuesto en los modelos teóricos.

Como resultado, se aprecia una pérdida de oportunidades para la investigación en este campo, sobre todo teniendo en cuenta que Irán experimentó un gran aumento en el uso de Internet y, con más de 45 millones de internautas, actualmente Irán es 13º país en el mundo con el mayor número de usuarios de Internet y el primer país en el oriente medio (Internet World Stats, 2013).

La lengua persa es el idioma oficial de Irán y uno de los idiomas oficiales de Afganistán. Se habla también en Tayikistán y partes de Uzbekistán. Hay más de cuatro millones de iraníes viviendo en los Estados Unidos, Canadá, Europa, Australia y en otras partes del mundo. Actualmente, la lengua persa se enseña en universidades e instituciones de Norteamérica y Europa. En los últimos años, aparecieron muchos sitios web que proporcionan informaciones en persa. En particular, la mayor parte de los periódicos y las revistas iraníes tienen sitios web oficiales con artículos diarios en persa.

El idioma persa es la lengua de muchos documentos publicados en la red y aproximadamente se utiliza por el 0,8% de todos los sitios web (W3Techs, 2015). El número de blogs en persa también ha experimentado un crecimiento espectacular, colocando a esta lengua entre los diez principales idiomas de la blogosfera mundial (Megerdoomian, 2008).

Debido a la naturaleza y características particulares de la lengua persa en comparación con otros idiomas como el inglés, el diseño de un sistema de RI en persa requiere consideraciones especiales. La recuperación de información en persa comenzó a desarrollarse en mayor medida, sobre todo, desde la llegada de Internet en la sociedad iraní. A partir de ahí, los investigadores se han dado cuenta de la necesidad de desarrollar herramientas que se adaptan a la lengua y a los documentos persas.

Cuando más materias y documentos persas están disponibles en la Web, es evidente que necesitan ser desarrollados más instrumentos de búsqueda para conseguir mejor acceso con fuentes de informaciones en ese idioma. Teniendo en cuenta estas circunstancias y dado el crecimiento que experimenta hoy en día los sistemas de RI en persa sobre todo en la Web, se propone abordar un estudio de la RI de los documentos

persas y la propuesta de algunos algoritmos y técnicas tendentes a mejorar la recuperación de información en persa.

1.3. Objetivos y Resultados Esperados

El propósito del trabajo desarrollado en la presente tesis doctoral es analizar la recuperación de información en persa y poner de manifiesto los factores que afectan a la eficiencia en los sistemas de RI cuando trabajan con documentos en persa. Para lograr estos objetivos, en primer lugar, se debe estudiar las características de la escritura de texto en persa desde el punto de vista de un sistema de RI. Al iniciar el análisis computacional del texto, se podría enfrentar una gran cantidad de ambigüedades por las características del idioma persa y su transcripción especial. Por lo tanto, se presentan los desafíos y problemas que la escritura y la lengua persa pueden tener desde la perspectiva de un sistema de RI textual. Como conclusión se plantea que el texto persa necesita de una pre-normalización o estandarización para cumplir con los criterios de un sistema de RI.

En segundo lugar, habría que analizar los trabajos ya realizados en el ámbito de RI para recoger las técnicas y algoritmos utilizados en diferentes etapas del proceso de recuperación de información. Este análisis permite deducir las debilidades de estos modelos que influyen en la precisión de los documentos recuperados por un sistema de RI. Por lo tanto, se han estudiado los trabajos previos realizados por diferentes grupos de investigadores para examinar detalladamente cada uno a fin de conocer las técnicas y algoritmos propuestos. El análisis realizado muestra que las técnicas utilizadas, sobre todos en la parte de la representación de los documentos, deben mejorar para aumentar la eficiencia de los sistemas de RI de documentos en persa.

En tercer lugar, se ha preguntado por qué los usuarios que utilizan la lengua persa en la Web no pueden conseguir documentos tan relevantes en comparación con los usuarios que utilizan, por ejemplo, el inglés. Para responder a esta pregunta se ha evaluado el rendimiento y calidad de los buscadores de web frente a los documentos en persa, particularmente en Google, dado que es utilizado por el 92% de los usuarios

iraníes. El motivo de esta evaluación es, sobre todo, conocer la estrategia de cómo Google analiza los documentos en persa para una búsqueda.

Analizando trabajos previos sobre este tema, se ha constatado que no hay ninguno que proponga la manera de identificar automáticamente las palabras vacías en un sistema de RI para documentos escritos en persa. Por lo tanto, la última aportación de este trabajo es desarrollar un método automático que permite identificar las palabras vacías para sistemas de RI en dichos documentos.

1.4. Estructura del Documento

Esta memoria se estructura en tres partes. En la primera, la parte teórica, se realiza una introducción al campo de RI, la lengua persa y sus características a considerar en un sistema de RI. La segunda parte, la parte experimental, es el núcleo de la memoria describiendo los trabajos desarrollados. Finalmente, el tercer bloque lo constituye una serie de apéndices en los que se presenta material que resulta de interés en el ámbito de la tesis. A continuación, presentamos un breve resumen del contenido de cada uno de los capítulos.

Parte I: Introducción y la Lengua Persa

Capítulo 1: En este capítulo se hace una breve introducción al campo de la recuperación de información. Tras presentar una serie de conceptos básicos se describe, sobre todo, la parte de representación de los documentos de un sistema de RI que depende de la lengua en la cual están escritos los documentos. Al mismo tiempo se exponen la motivación de realizar esta tesis y también los objetivos y resultados esperados por los experimentos desarrollados durante este trabajo; además, se describe brevemente la estructura del presente documento.

Capítulo 2: Se dedica a la visión general de la lengua persa, su alfabeto y su escritura. Se describe también las características específicas de la lengua persa y plantea los

principales retos en el procesamiento del texto y la complejidad asociada a la lengua frente a un sistema de RI textual.

Parte II: Recuperación de Información en Persa

Capítulo 3: Es un análisis de los trabajos previos y los últimos avances en el ámbito de la RI de los documentos persas. Describimos los más importantes, especialmente todos aquellos que se refieren a la parte dependiente del lenguaje en la RI y que han servido como base a la investigación.

Capítulo 4: Se analiza la tarea de RI de los documentos persas en Internet. Evaluamos la recuperación de información en persa por el motor de búsqueda Google, el buscador más utilizado por los usuarios iraníes como una herramienta de RI en la Web.

Capítulo 5: Se dedica a la explicación de una metodología automática para construir una lista de palabras vacías para sistemas de RI en documentos persas. Se explica también la contribución que estas palabras pueden tener al nivel de la reducción del tamaño de índice.

Capítulo 6: Este último capítulo recoge las conclusiones y aportaciones de este trabajo de tesis, así como las vías de desarrollo futuro del mismo.

Bibliografía: Se recopilan las referencias bibliográficas que son recursos utilizados para la realización de este documento.

Parte III: Apéndices

Apéndice A: Se presentan las publicaciones y otros resultados obtenidos que se pretenden publicar en el futuro.

Apéndice B: En esta sección se encuentran todos los códigos de fuente que hemos utilizado para desarrollar nuestros experimentos (ver la Sección 4.7.3).

Apéndice C: Es un artículo escrito en la lengua persa en el que verificamos la ley de Zipf utilizado en el capítulo 5 (ver la Sección 5.3.2.1).

Capítulo 2

La Lengua Persa y sus Retos en un Sistema de RI

Resumen

La primera parte de este capítulo se dedica a la visión general de la lengua persa con un énfasis en algunos aspectos de análisis computacional del texto. En la segunda parte se plantean los principales desafíos y problemas de la lengua persa al nivel de procesamiento del texto en un sistema de recuperación de información. Debido a la ambigüedad del texto escrito y variedad en la escritura de palabras, los documentos persas necesitan una pre-normalización o preparación del texto para los sistemas de recuperación de información. Además, la morfología compleja, las diferentes formas de una letra y el uso o la eliminación del espacio entres los caracteres de una palabra crean obstáculos en el proceso de tokenización y lematización del texto persa. La representación de los documentos persas en un sistema de recuperación de información requiere procesamientos específicos para determinar los términos de índice que corresponden a un documento.

2.1. Introducción

El persa, también conocido como Farsi[1] (فارسی) o Parsi (پارسی), es la lengua oficial de Irán. Es también uno de los dos idiomas principales hablado en Afganistán y la lengua principal en Tayikistán, hablado también en unas partes de Uzbekistán, Omán, Emiratos Árabes e incluso en India [Figura 2.1] sin contar con los millones de persas que hay en la diáspora. Los ambientes locales han influenciado al persa hablado en estos países. Esto es especialmente verdad en Tayikistán puesto que fue aislado de otros países de lengua persa durante la era soviética. El persa en este país tiene muchos

[1] La palabra "farsi" fue originalmente la forma arábiga para expresar "parsi", el antiguo nombre del idioma, debido a la carencia del fonema /p/ en árabe.

préstamos rusos y se utiliza el alfabeto ruso para escribir las palabras persas. El persa pertenece a la familia de lenguas indoeuropeas de lo contrario a lo que la mayoría piensa al creer que es una lengua árabe.

La lengua descrita en este documento es principalmente el persa hablado en Irán. Este capítulo nos da una visión general de la lengua persa con un énfasis en algunos aspectos interesantes para un análisis computacional del texto escrito en esta lengua.

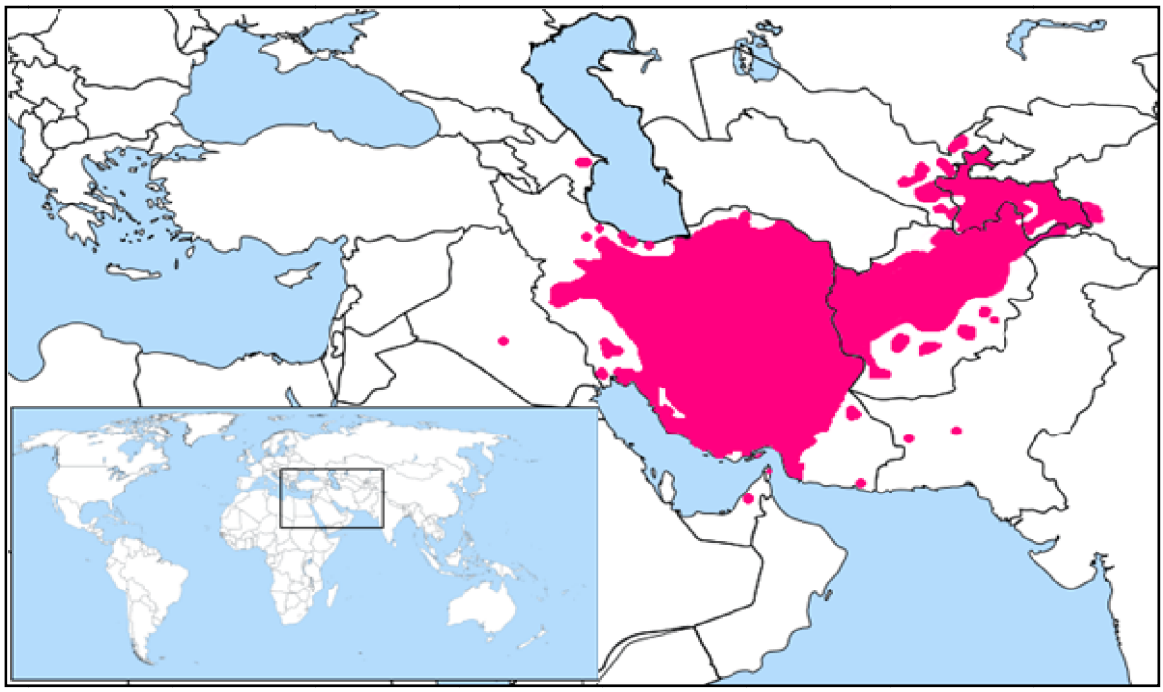

Figura 2.1: Extensión de la lengua persa, tomado de (Wikipedia, 2012)

2.2. Origen

El persa es una de las lenguas más antiguas del mundo. La lengua persa, dentro de la familia indoeuropea, pertenece a las lenguas indo-iraníes que a su vez se dividen en lenguas iraníes e indo-arias (Amtrup, et al., 2000). El persa es una lengua iraní noroccidental y está documentado con diferentes variantes desde hace más de 25 siglos. Naturalmente en todo el período documentado, la lengua ha cambiado notablemente por lo que el persa más antiguo es una lengua ininteligible para un hablante de persa moderno. La evolución cronológica de la lengua persa es la siguiente (Mazdak, 2004):

- **Persa antiguo** o *Persa Aqueménida*, documentado en inscripciones cuneiformes esculpidas durante el Imperio *Aqueménida*, hasta aproximadamente el 300 a. C.
- **Persa medio** (*Pahlavi* o persa *sasánida*), documentado especialmente durante el imperio de los sasánidas y coetáneo del idioma parto (iranio suroccidental).
- **Persa moderno** comienza alrededor del año 900 de la era cristiana hasta nuestros días y es desde esa época cuando se van formando las actuales tres grandes lenguas persas: el persa contemporáneo, el *darí* y el *tayico*.

La lengua en sí ha evolucionado enormemente a lo largo del tiempo, habiendo cambiado tanto en el nivel fonológico como en el morfosintáctico. Además, debido al desarrollo tecnológico y el contacto con otros pueblos de oriente medio, el persa presenta un buen número de préstamos léxicos procedentes de otras lenguas, situación que también se da en el resto de lenguas de la región (Wikipedia, 2014).

2.3. Sistema de Escritura

En su forma más antigua, la lengua persa aparece en las inscripciones cuneiformes (estilo pictograma) ya en el siglo VI a. C. Más tarde, los persas inventaron un nuevo alfabeto llamado "*Pahlavi*" para reemplazar el alfabeto cuneiforme. Sin embargo, después de la conquista árabe en 651, los persas adoptaron la escritura unificada árabe.

Tabla 2.1: Letras persas que no existen en árabe

Sonido	Forma	Nombre Unicode	Unicode
/p/	پ	Peh	067E
/ʧ/ (ch)	چ	Tcheh	0686
/ʒ/ (zh)	ژ	Jeh	0698
/g/	گ	Gaf	06AF

La escritura moderna de persa utiliza el alfabeto árabe, pero con la adición de cuatro letras que no aparecen en árabe. Estas son: 'پ', 'چ', 'ژ' y 'گ' (ver la Tabla 2.1).

En lo que sigue vamos a usar los corchetes [ ] para representar la pronunciación en español de una letra o palabra persa y las barras / / para representar la transcripción fonética de un sonido.

Como todos los sonidos del alfabeto árabe no existen en la lengua persa entonces más de una letra puede representar el mismo sonido. Por ejemplo, hay cuatro letras en persa para el sonido /z/ (ز , ذ , ض , ظ) y tres para el sonido /s/ (ث , س , ص). Además, un sonido concreto puede ser representado por varios símbolos contribuyendo a aumentar la confusión. El alfabeto persa tiene 32 letras y se escribe de la derecha a izquierda. Una lista completa del alfabeto se presenta en la Tabla 2.2 de la misma página y la siguiente.

Tabla 2.2: Alfabeto persa

Letra	Sonido	Nombre Unicode	Unicode	Como en
آ	/ɒ/	alef con madda arriba	0622	ámbito
ا	/a/	alef	0627	atención
ب	/b/	beh	0628	barco
پ	/p/	peh	067E	pez
ت	/t/	teh	062A	té
ث	/s/	theh	062B	sol
ج	/ʤ/	jeem	062C	la j inglesa: Jack
چ	/ʧ/	tcheh	0686	Chino
ح	/h/	hah	062D	la h inglesa: House
خ	/x/	khah	062E	jueves
د	/d/	dal	062F	dos
ذ	/z/	thal	0630	la z inglesa: zoo
ر	/r/	reh	0631	real
ز	/z/	zain	0632	lLa z inglesa: zoo
ژ	/ʒ/	jeh	0698	la j francesa: jeudi
س	/s/	seen	0633	sol
ش	/ʃ/	sheen	0634	sh en inglés: show
ص	/s/	sad	0635	sol
ض	/z/	dad	0636	la z inglesa: zoo
ط	/t/	tah	0637	té

Tabla 2.2 (continuación)

Letra	Sonido	Nombre Unicode	Unicode	Como en
ظ	/z/	zah	0638	la z inglesa: zoo
ع	/ʔ/	ain	0639	oclusiva glotal[2]
غ	/ɣ/	ghain	063A	una ¨q¨ gutural
ف	/f/	feh	0641	forma
ق	/ɣ/	qaf	0642	una ¨q¨ gutural
ک	/k/	kaf	06A9	kilo
گ	/g/	gaf	06AF	gordo
ل	/l/	lam	0644	luz
م	/m/	meem	0645	mesa
ن	/n/	noon	0646	norte
و	/v, u:/	waw	0648	la w inglesa: work
ه	/h/	heh	0647	la h inglesa (suave)
ی	/i:, y/	yeh	06CC	Irán , rey

2.3.1. Las Reglas Principales de la Escritura

Las reglas de escritura son las siguientes:

- Los caracteres en persa toman una de las cuatro formas: inicial, medial, final y aislada dependiendo del lugar donde se ocurren en la secuencia de texto. La forma inicial indica que ningún carácter está agregado a la derecha (es decir, no hay agregación del carácter antes de él, sino que hay uno que sigue al carácter). Los caracteres son en forma intermedia si tienen un carácter agregado tanto antes como después de ellos. El último carácter en una palabra marca el final de la palabra y se escribe de la forma final o aislada. Por ejemplo en la Tabla 2.3 podemos ver las diferentes formas de dos sonido ‘B’ y ‘G’.

[2] La oclusiva glotal es un sonido que se articula por un cierre momentáneo completo de la glotis en la parte posterior de la garganta, como en el inicio repentino de una vocal.

Tabla 2.3: Formas de los caracteres en persa

Sonido	Final	Medial	Inicial	Aislada
/B/	ـب	ـبـ	بـ	ب
/G/	ـگ	ـگـ	گـ	گ

- Las letras en una palabra están conectadas entre sí a excepción de las letras ا, آ, د, ذ, ر, ز, ژ, و. Estas letras, debido a su forma, no pueden ser escritas de modo que conecten con el carácter siguiente en una palabra.
- Las vocales pueden ser cortas o largas. Las cortas son [a], [e], [o] y se representan con un signo "vocálico" encima o debajo de la consonante a la que vocalizan (zebar ـَ /æ/, zir ـِ /e/ , pish ـُ /o/). Por lo general, las vocales cortas no se escriben.
- Las vocales largas [ā], [i], [u] aparecen como una letra propiamente dicha, a continuación de la consonante a la que acompañan (alef آ /a/, ye ی /i:/, waw و /v/).
- La vocal larga [a] se representa por "آ" en la posición inicial y por "ا" de otro modo (ver la Tabla 2.4).

Tabla 2.4: La vocal larga [a]

Palabra	Pronunciación	Traducción
آب	[ab]	agua
باد	[bad]	Viento

2.3.2. Los Números

Los números persas tienen el mismo origen que los números latinos y se escriben y se leen de la izquierda a la derecha (ver Tabla 2.5).

Tabla 2.5: Los números en la lengua persa

Números	En español	Unicode	Números	En español	Unicode
۰	0	06F0	۵	5	06F5
۱	1	06F1	۶	6	06F6
۲	2	06F2	۷	7	06F7
۳	3	06F3	۸	8	06F8
٤	4	06F4	۹	9	06F9

2.3.3. Los Límites de la Palabra

En el texto persa, las palabras son separadas generalmente por un espacio, un signo de puntuación y por las distintas formas que los caracteres pueden tener en función de su posición en la palabra.

2.3.3.1. Espacio

Los límites de palabras son denotados generalmente por espacios. Sin embargo, las palabras compuestas, las construcciones ligeras del verbo y los morfemas desmontables (es decir, morfemas que siguen una palabra terminando en un carácter de la forma final) pueden aparecer sin un espacio que los separe.

2.3.3.2. Puntuación

Ciertos signos de puntuación indican límites de la frase. El punto (.) marca el límite de una oración pero también puede aparecer en la formación de abreviaturas o acrónimos. Los signos de exclamación e interrogación son indicadores de límites inequívocos. Aparte de la barra (/) que se utiliza en los números y el guión (-) que podría ser utilizado para separar palabras compuestas, los otros signos de puntuación indican

inequívocamente el límite de las palabras. Estos incluyen la coma, las comillas, el punto y coma, paréntesis y dos puntos.

2.3.3.3. La Forma del Carácter

Un carácter de forma final indica el final de una palabra y puede ser utilizado para determinar el límite de una palabra. Por lo tanto, dos palabras concatenadas se pueden poner en palabras separadas si la primera palabra termina en un carácter de forma final. Pero si la primera palabra termina en uno de los caracteres que tienen solamente una forma entonces el final de la palabra no está claro.

2.4. Nombres

2.4.1. Género

La lengua persa no tiene género gramatical. El sexo se indica por medios léxicos, por ejemplo گاو ماده [gav e madde] que se traduce “toro hembra” (la vaca, en español).

2.4.2. Genitivo

El genitivo es el caso de complemento nominal. El complemento nominal especifica la palabra (o el grupo de palabras) a que se relaciona. En persa, este caso se introduce por [e] o [ye] que es una vocal corta y no se escribe کتاب سارا [ketab e Sara] (el libro de Sara, en español).

2.4.3. Sustantivo

El sustantivo tiene dos formas: singular y plural. El plural se indica con los diferentes sufijos que pueden ser adjuntos o separados de los sustantivos. La mayoría de los

sustantivos forman su plural añadiendo el sufijo ها [ha] al singular. Sin embargo, la forma agregada no puede aparecer en las palabras que se terminan en ه [he] (ver Tabla 2.6).

Tabla 2.6: Morfema plural ها [ha]

Palabra singular	Morfema plural	Forma adjunta	Forma separada	Traducción
عکس	ها [ha]	عکسها	عکس ها	foto(s)
آینه	ها [ha]	-	آینه ها	espejo(s)

Existe también el sufijo ان [an] que se usa típicamente para indicar los sustantivos de ser humano o animal. Si cualquier sustantivo termina en un ا [a] o و [v], le sufijo [an] se convierte en یان [yan]. Hay muchas excepciones donde la palabra se termina en [an] pero no es una forma plural, por ejemplo, la palabra قهرمان [ghahraman] (héroe, en español). En el caso que un sustantivo se termina en un ه [he] antes de añadir el sufijo, ه [he] se convierte en گ [g]. La Tabla 2.7 muestra los diferentes casos del signo plural [an] en una palabra.

Tabla 2.7: Morfema plural ان [an]

Palabra	Forma adjunta	Forma separada	Traducción
مرد	مردان	-	hombre(s)
دانا	دانایان	-	sabio(s) (persona)
دانشجو	دانشجویان	-	estudiante(es)
پرنده	پرندگان	-	pájaro(s)

La forma plural de algunos sustantivos se hacen añadiendo signos plurales árabes como ون [vn], ین [in] y ات [at]. Pero si un sustantivo termina en un [a], [v], [eh] o [ye] el sufijo [at] se convierte en [jat]. Sin embargo, hay también muchas palabras que se terminan con estos signos y no son palabras plurales (ver Tabla 2.8).

Tabla 2.8: Morfemas plurales árabes en el texto persa

Palabra singular	Morfema plural	Forma adjunta	Forma separada	Traducción
انقلابی	ون [vn]	انقلابیون	-	revolucionario(s)
مسافر	ین [in]	مسافرین	-	pasajero(s)
انتخاب	ات [at]	انتخابات	-	elección(es)

Por otra parte hay algunos sustantivos adoptados de la lengua árabe que tienen formas plurales irregulares y se utilizan también persa (ver Tabla 2.9).

Tabla 2.9: Plurales irregulares

Palabra singular	Palabra plural	Traducción
کتاب	کتب	libro(s)

2.4.4. Artículos

No existen artículos definidos o determinados por lo que un sustantivo es determinado por sí mismo. El artículo indeterminado es la letra ی [i,y] que viene después del sustantivo, por ejemple پسری [persar i] (un chico, en español).

2.4.5. Adjetivo

Contrariamente al español, el adjetivo es invariable y no concuerda en género o número (singular o plural) con el sustantivo que modifica. En persa, al igual que en español, el adjetivo va detrás del sustantivo pero el adjetivo se relaciona al sustantivo por la preposición del genitivo (ver Tabla 2.10).

Tabla 2.10: Adjetivos en persa

Adjetivo singular	Pronunciación	Traducción	Adjetivo plural	Pronunciación	Traducción
خانه بزرگ	Khane ye bozorg	la casa grande	خانه های بزرگ	Khaneha ye bozorg	las casas grandes
پیراهن سفید	Pirahan e sefid	la camiseta blanca	پیراهن های سفید	Pirahanha ye sefid	las camisetas blancas

2.4.6. Morfemas de Comparación

Los morfemas que indican comparación aparecen añadidos a los adjetivos. Se pueden también adjuntar a los adverbios de modo. Los signos de comparación son comparativos o superlativos y pueden aparecer unidos a la palabra o en forma individual. El comparativo se forma añadiendo el sufijo تر [tar] al adjetivo y el complemento del comparativo se introduce por la preposición از [az] (de, en español) y normalmente, se coloca después del comparativo. El superlativo se forma agregando el sufijo ترین [tarin] al adjetivo. El problema de distinguir estos sufijos siempre surge cuando se unen a las palabras (ver Tabla 2.11).

Tabla 2.11: Signos de comparativos

Comparación	Morfema	Forma adjunta	Forma separada	Traducción
Comparativo	تر [tar]	بزرگتر از	بزرگ تر از	más grande que…
Superlativo	ترین [tarin]	آسانترین	آسان ترین	lo más fácil

2.4.7. Pronombres Personales

Los pronombres personales tienen dos formas, libre y ligada. Los pronombres libres son independientes y pueden aparecer a solas en una oración. Los ligados no se usan independientemente y siempre aparecen acompañados y agregados a alguna otra palabra (verbo, sustantivo…). Los pronombres personales ligados se usan siempre con otras palabras. El español no tiene un equivalente a ellos. Según la palabra que acompañan, estos pronombres toman diferentes funciones. Con un sustantivo,

funcionan como adjetivos posesivos, por ejemplo پدرم [pedar-am] (mi padre, en español) y con un verbo, funcionan como pronombres de objeto, por ejemplo می بینمش [mibinam-ash] (la (lo) veo, en español). Hay muchas excepciones donde una palabra se termina con los signos de pronombres personales ligados pero no tiene la función de un pronombre personal.

2.5. Verbos

Los verbos del idioma persa se clasifican en dos grupos principales: simples y compuestos. Un verbo simple consta de un solo verbo. Por ejemplo داشتن [dashtan] (tener, en español). Un verbo compuesto consta de una o más palabras y un verbo simple. Una peculiaridad del persa es que no tiene muchos verbos simples (a diferencia del español). La mayoría de los verbos son compuestos. Los verbos compuestos consisten en elementos pre-verbales (que pueden ser sustantivos, preposiciones o adjetivos) seguidos por un verbo simple tal como los verbos کردن [kardan] (hacer, en español), دادن [dâdan] (dar) o زدن [zadan] (pegar). En esta estructura, conocido como construcción de verbos ligeros, el verbo pierde su significado original pero se asocia con otros elementos para formar un nuevo verbo. El significado de un verbo ligero es no composicional; es decir, no puede ser obtenido traduciendo cada elemento por separado. Por ejemplo, en la construcción del verbo ligero به دنیا آمدن [be donya amadan] (nacer, en español) si traducimos palabra por palabra tenemos "al mundo venir".

Un verbo simple se divide en dos grupos de acuerdo con la raíz que utiliza en su formación. Cada verbo tiene dos raíces, la raíz del presente y la raíz del pasado. No hay reglas regulares para obtener la raíz del presente y se debe especificar en el léxico, mientras que la raíz del pasado se deriva fácilmente desde el infinitivo del verbo. Además de la raíz del verbo, los siguientes elementos también participan en la formación del sistema de inflexión verbal (Megerdoomian, 2004) :

- **Prefijos**: El prefijo imperfectivo می [mi] y el morfema ب [b] o بی [bi] que caracterizan el subjuntivo y el imperativo. La negación del verbo está marcada por ن [n] o نی [ni].

- **Inflexiones personales**: Las inflexiones personales del presente, pasado y imperativo se utilizan en la conjugación del verbo. Todas las formas verbales concuerdan en número y persona.
- **Sufijos**: El sufijo نده [ande] (a es una vocal corta que no se escribe) se utiliza para el participio presente y ه [he] se utiliza para formar el participio pasado.
- **Morfema de causales**: Las causales se obtienen por el afijo ان [an] o انی [ani] al final de la raíz del presente de un verbo. Las inflexiones personales y sufijos pueden ser adjuntos a la raíz del presente causativo para derivar todas las formas verbales de la construcción causativa.
- **Auxiliares**: La conjugación de verbos utiliza un número de auxiliares en las formas compuestas. La forma enclítica[3] del verbo auxiliar بودن [budan] (ser, en español) se utiliza en la formación perfecta de todos los verbos. El verbo خواستن [khastan] (querer, en español) se utiliza como un auxiliar en la formación de los tiempos de futuro. El verbo شدن [shodan] (convertirse en, en español) forma las construcciones pasivas.

El sistema inflexión completa se puede conseguir por las diversas combinaciones de estos elementos.

2.6. Desafíos y Problemas de la Lengua Persa en un Sistema de RI

La tecnología de la información, en su dominio de actividad, tiene relación con documentos. Un documento es un objeto de datos, de naturaleza textual generalmente, aunque la evolución tecnológica ha propiciado la profusión de documentos multimedia, incorporándose al texto fotografías, ilustraciones gráficas, vídeo animado, audio, etc. (Castillo Sequera, 2010). El lenguaje natural, entendido como la herramienta que utilizan las personas para expresarse, tiene propiedades específicas que reducen la efectividad de los sistemas de la RI textual. Estas propiedades son la variación y la

[3] Partícula o parte de la oración que se liga con el vocablo precedente, formando con él una sola palabra, como en el español los pronombres pospuestos al verbo.

ambigüedad lingüística (Vallez, et al., 2007). La variación lingüística se refiere a la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ambigüedad lingüística se produce cuando una palabra o frase permite más de una interpretación. Ambos fenómenos inciden en el proceso de RI aunque de forma distinta. La variación lingüística provoca el silencio documental, es decir la omisión de documentos relevantes para cubrir la necesidad de información porque no se han utilizado los mismos términos entre la consulta y los documentos relevantes. En cambio, la ambigüedad lingüística implica el ruido documental, es decir la inclusión de documentos que no son significativos, ya que se recuperan también documentos que utilizan el término pero con significado diferente al requerido. Estas dos características dificultan considerablemente el tratamiento automatizado del lenguaje y por secuencia dificultan el proceso de la RI (Baeza-Yates, 2004).

La mayor parte de los modelos y técnicas empleados en RI utilizan en algún momento recuentos de frecuencias de los términos que aparecen en los documentos y en las consultas. Esto implica la necesidad de normalizar dichos términos, de manera que los recuentos puedan efectuarse de manera adecuada. Pero la lengua persa debido a sus ambigüedades en el texto escrito, la dispersión en posición alfabética diferente, la ambigüedad del límite de la palabra y la variedad de ortografía pueden causar problemas en el procesamiento del texto y por secuencia en el proceso de búsqueda de información (Mortezai, 2006). Estos problemas alteran los recuentos de frecuencias de los términos en el documento y en secuencia, afectan a la tasa de recuperación en un sistema de RI. Antes de cualquier proceso de RI, el texto persa necesita experimentar una pre-normalización que es una forma de estandarización del texto y que reduce la variedad en favor de la uniformidad. En efecto, la pre-normalización es establecer reglas que aseguren la interconexión de sistemas y que faciliten el tratamiento y la transferencia de información (García Gutiérrez, 1985).

La lengua persa es uno de los idiomas con menos recursos y menos estudiados desde el punto de vista computacional. A continuación, describimos, en primer lugar, algunas características específicas de la lengua persa que son diferentes de otros idiomas como el español y pueden ser fuente de problemas en el procesamiento del

texto. Después, exponemos los principales desafíos y problemas abiertos en el procesamiento del texto relacionado con la tarea de RI.

2.6.1. Características de la Lengua Persa

En esta sección explicamos algunas características de la lengua persa que diferencian con otros idiomas como el español. Estos detalles nos ayudan a comprender por qué las herramientas disponibles del procesamiento de texto, que son sobre todo para los idiomas anglosajones y latinos, no pueden adaptarse para otros idiomas de naturalezas diferentes como la lengua persa.

- Como la escritura es de la derecha a la izquierda entonces, hay una discordancia entre el texto normal y los textos que contienen las formulas matemáticas, químicas, las notas musicales y también los signos gráficos que se utilizan por todo el mundo y se lee de la izquierda a la derecha.
- La lengua Persa es una lengua *pro-drop* (de la palabra inglés "*pronoun-dropping*") con el orden de palabras canónico SOV (Sujeto-Objeto-Verbo). Una lengua con omisión de sujeto (lengua "*pro-drop*") no requiere la aparición de un sujeto sintáctico explícito como sucede con el español. En cambio el francés o el inglés no son lenguas *pro-drop,* ya que requieren obligatoriamente un sintagma nominal como sujeto o bien un pronombre personal en posición de sujeto (Mahootian, 1997). Pero la lengua persa está considerada como un idioma con el orden libre de palabras debido a muchas excepciones en el orden de palabras en sus textos (Mahootian, 1997).
- Los verbos están marcados por tiempo y aspecto[4] correspondiendo con el sujeto en persona y número. Pero el número correspondiente del verbo puede ser ignorado para sujetos inanimados plurales o para respeto de los sujetos

[4] El aspecto verbal es una propiedad lingüística característica de los verbos, así como de las perífrasis verbales, que sirve para señalar si la acción que expresan ha concluido en el momento indicado en la oración, o si aún no lo ha hecho.

singulares. Por ejemplo, “el rector han dicho….” es una forma de respeto al rector.

- Cada verbo se conjuga en su propio tiempo. Por ejemplo, cuando en español se dice “quiero leer mi libro “, el segundo verbo se utiliza en su forma infinitiva mientras que en persa los dos verbos se conjugan. Además, los verbos conjugados son diferentes para cada persona.
- Por lo general, ninguna vocal corta se escribe en una frase así que el hecho de tener palabras que se parecen pero la pronunciación es diferente (homógrafos) o tienen significados diferentes (homónimos) es una ambigüedad muy popular en la lengua persa.
- A pesar de que los verbos se encuentran al final de la oración, por lo demás la lengua persa es principalmente de cabeza-inicial (Mahootian, 1997). Es decir, los verbos por lo general se localizan en el final de la frase después de sus argumentos y objetos, y la cabeza de un sintagma nominal aparece antes de sus modificadores.
- No hay reglas para forzar sustantivos incontables en singular. Incluso las palabras que son incontables pueden aparecer en forma plural.
- El idioma persa es una lengua derivativa y generativa en la que muchas palabras nuevas se pueden construir mediante la concatenación de palabras y afijos. Así que la posibilidad de encontrar una nueva palabra en el texto que no está disponible en el léxico es muy probable.
- Las palabras y frases pueden ser omitidas en una sentencia de acuerdo con una simetría sintáctica o semántica. La omisión del sujeto es también muy frecuente en las oraciones. En este caso, la persona y el número incrustados en el verbo pueden desempeñar el papel de sujeto.
- En muchos casos, los adjetivos se pueden insertar en lugar de los nombres sin ningún cambio léxico y esto puede causar ambigüedades estructurales o semánticas en sintagmas nominales.

2.6.2. Desafíos y Problemas de la Escritura

El idioma persa es uno de los idiomas con tareas complejas y desafiantes del procesamiento de texto (Shamsfard, 2011). Al tener diferentes formas de escritura, el principal problema es que necesita una unificación o pre-normalización del texto para convertir todas las formas de escritura en un único estándar antes de procesar el texto. A continuación, se enumeran algunos problemas y desafíos que deben ser abordados en estas unificaciones.

2.6.2.1. Ambigüedades en el Texto Escrito (Homógrafo y Homónimo)

Ciertas ambigüedades se presentan en un análisis computacional del texto puesto que la misma forma de superficie puede representar diversos morfemas. Además, las vocales cortas no están marcadas en el texto escrito y eso da lugar a diversas posibilidades de análisis. Las vocales cortas son: " ́/a/, ُ/o/, ِ /e/ ". El ejemplo siguiente en la Tabla 2.12 ilustra bien estas ambigüedades. La palabra escrita como مردم "*mrdm*" se puede pronunciar con vocales /a/ o bien /o/ o las dos pero tiene diferentes significados.

La ambigüedad se puede también dar en el caso de "*Tashdid*". *Tashdid* es un signo que se coloca sobre una letra en una palabra para duplicar su pronunciación y su símbolo es " ّ ". Su efecto equivale en español a usar una letra en doble en una palabra. Esta letra se pronuncia dos veces; el primero sin vocal y la segunda vez con el sonido de la vocal. *Tashdid* se puede escribir u omitir en el texto escrito, así como las vocales cortas.

Tabla 2.12: Ambigüedad del texto persa

Palabra escrita	Pronunciación	Traducción
مردم	mardom	gente
مردم	mordam	morí
مردم	mardam	soy hombre

2.6.2.2. Supresión de Caracteres

Existen algunos caracteres que se suprimen en algunas palabras escritas. Por ejemplo, en los dos casos siguientes mostrados en la Tabla 2.13 el carácter "ا" está eliminado en la palabra pero se pronuncia.

Tabla 2.13: Supresión del carácter en una palabra

Palabra escrita	Pronunciación	Traducción
اسحق	Eshagh	Isaac (nombre)
اسمعیل	Esmail	Ismael (nombre)

2.6.2.3. Diferentes Morfemas para el Mismo Sonido

Un sonido se puede escribir con diferentes caracteres. El sonido /s/ se puede escribir con "س"; "ص" o "ث", el sonido /z/ con "ز"; "ذ"; "ض" o "ظ", y el sonido /t/ con "ت" o "ط". (Véase Tabla 2.14).

Tabla 2.14: Diferentes morfemas para el mismo sonido

Palabra escrita	Pronunciación	Traducción
سیمان	siman	cemento
صدا	seda	sonido
ثابت	sabet	fijo
مذهب	mazhab	religión
زیبا	ziba	bonito
ضمیمه	zamimeh	apéndice
ظاهر	zaher	aparente
تنهایی	tanhaii	soledad
مطلب	matlab	sujeto

2.6.2.4. Los Diversos Puntos en una Letra

Los diversos puntos encima o abajo de los caracteres pueden causar errores para leer una letra en una palabra. La importancia de puntos (la posición y el número de puntos en una letra) puede crear problemas en la identificación óptica de los caracteres. Por ejemplo, la diferencia entre "ر" , "ز" y "ژ" ; entre "د" y "ذ' ; entre "ب" , "ت" , "ث" y "پ" o entre "خ" , "ح" , "ج" y "چ" solo se reside en puntos. Véanse las palabras siguientes de la Tabla 2.15 donde el significado de la palabra "بر" cambia sólo cambiando los puntos.

Tabla 2.15: Múltiplos puntos en una palabra

Palabra escrita	Pronunciación	Traducción
بر	bar	en
پر	par	pluma
تر	tar	mojado
پز	poz	postura
بز	boz	cabra
تز	tez	tesis

2.6.2.5. Diversos Equivalentes para los Términos Científicos

Resulta común que los especialistas utilicen diferentes términos para referirse al mismo concepto. Los términos científicos que vienen sobre todo de la lengua inglesa pueden tener diferentes equivalentes dependiendo de dónde se utilicen. Por ejemplo, podemos ver que los bibliotecario y los especialistas de la tecnología de información utilizan seis equivalentes para el término "*Manual*", nueve equivalentes para el término "*Online*", doce equivalentes para el término "*Layout*" y trece equivalentes para el término "*Cross reference*". De estos equivalentes existen en todas las partes de la ciencia (Mortezai, 2006).

2.6.2.6. Variedad de Transcripción para los Términos Extranjeros

No hay reglas específicas para escribir (selección de letras) y traducir nombres de personas, organizaciones, sustancias y compuestos químicos, herramientas y equipamientos, lugares geográficos, etc. provenientes de otros idiomas. Cada experto (escritor o traductor) según su gusto, intuición, conocimiento de la lengua original y de la suya propia puede elegir una transcripción distinta para los nombres extranjeros utilizando en los textos persas. Esta carencia de coordinación se observa incluso en las publicaciones de organizaciones científicas y culturales del país. Como estas palabras no están en el léxico persa (Véase ejemplos en la Tabla 2.16), entonces la tokenización y la corrección ortográfica no son fáciles (Shamsfard, 2011).

Tabla 2.16: Diferentes transcripciones para las palabras extranjeras

Palabra	Forma 1	Forma 2	Forma 3	Forma 4
Robinsón	رابینسون	روبینسون	ربینسون	روبنسن
Potasio	پتاسیم	پتاسیوم	پوتاسیم	پوتاسیوم
FID[5]	اف.آی.دی	فید		
Hidrógeno	هیدروژن	ئیدروژن		

2.6.2.7. Ortografía Continua o Separada

Ortografía de la lengua persa es tal que podemos escribir las palabras en forma continua y separada. La diversidad de las mismas palabras, que es en realidad la variación lingüística, provoca el silencio documental, es decir la omisión de documentos relevantes para una dada consulta en un sistema de RI. La posición del morfema plural, que se puede escribir en forma continua y partida, producirá el mismo efecto en los índices de un texto (Ashouri, 1996). Por ejemplo, en la Tabla 2.17 podemos ver algunas palabras con formas escritas diferentes.

[5] Fondation Internationale pour le Développement

Tabla 2.17: Diferentes ortografías para la misma palabra

Palabra	Forma 1	Forma 2	Forma 3	Forma 4
Montañas	کوه ها	کوهها		
Ali Reza (nombre)	علی رضا	علیرضا		
Limpieza en seco	خشکشویی	خشک شویی		
Caldera	آب گرم کن	آب گرمکن	آبگرم کن	آبگرمکن

2.6.2.8. Diversidad de Formas Plurales

Diversas formas plurales como (ان , ات , ها , ین , ون) y también formas plurales irregulares en persa pueden causar problemas para las bases de datos que utilizan palabras claves plurales. El usuario, al momento de buscar información, debe considerar todas las formas plurales para las palabras claves o bien, con el uso de métodos convencionales, truncar la palabra. En ambos casos existe la posibilidad de no cubrir algunas formas de palabras plurales irregulares (Véase la Tabla 2.18).

Tabla 2.18: Diferentes formas de plural

Palabra	Forma plural 1	Forma plural 2	Forma plural 3	Traducción
استاد	اساتید	استادان	استادها	profesor(es)
مدرسه	مدارس	مدرسه ها		escuela(s)
محقق	محققان	محققین		investigador(es)
مشکل	مشکلات	مشکلها	مشکل ها	dificultad(es)
معلم	معلمین	معلمان	معلم ها	institor(es)

2.6.2.9. Letras Importadas de la Lengua Árabe

Existen algunos sonidos importados del árabe como "*Tanwin*[6]" y "*Hamza*[7]" que se utilizan también en algunas palabras persas. Estas palabras se pueden escribir en otras formas diferentes. El principal problema de estos sonidos es que pueden ser escritos o ignorados en una palabra. En el caso que los sonidos están escritos, esto puede hacerse de varias formas. Por ejemplo las palabras "مسئله", "مسأله" y "مساله" son todas las formas para escribir la palabra "problema". Las dos primeras formas muestran la palabra escrita con *hamza* y la última muestra la palabra escrita sin *hamza* (sustituyendo de *hamza* por el carácter *Alef*). El mismo problema se produce para el sonido *tanwin* que normalmente está adjunto al carácter *Alef* y se ignora escribiéndolo. Por ejemplo, las dos palabras "حتما" y "حتماً" significan "ciertamente" pero la última palabra tiene *tanwin* mientras que la primera lo ignora (ver Tabla 2.19).

Tabla 2.19: Diferentes clases de escritura

Palabra	Forma 1	Forma 2	Forma 3
Problema	مسئله	مسأله	مساله
Responsabilidad	مسئوليت	مسؤليت	
Ciertamente	حتما	حتماً	
Inicio	ابتدا	ابتداء	
Otoño	پاييز	پائيز	

2.6.2.10. Espacio Adicional en una Palabra

En el texto persa, el espacio blanco no determina necesariamente el límite de una palabra. Puede aparecer dentro de una palabra o entre varias palabras. Por otro lado, puede que no haya un espacio entre dos palabras, especialmente cuando el ultimo

[6] El tanwin es un sonido "n" añade al final de la palabra, en determinadas circunstancias, por lo general funciona igual que la "a" y "an" en inglés.

[7] Un signo en la ortografía árabe usada para representar el sonido de una parada glótica, transcrito en inglés como un apóstrofe

carácter de la primera palabra tiene sola una forma de escribirse (la forma aislada). En estas situaciones la lengua persa puede ser similar a algunos idiomas asiáticos como el chino sin espacio entre palabras. Hay muchas palabras que se pueden escribir con el espacio, el espacio corto o sin espacio. El espacio corto es un pseudo-espacio (*zero-width-joiner* en la codificación de carácter Unicode) dentro de las palabras. La falta de un conjunto de reglas para la escriture permite escribir las mismas palabras y frases en múltiples formas por escritores nativos persas (Véase la Tabla 2.20).

Tabla 2.20: Espacio adicional en una palabra

Palabra	Forma 1	Forma 2	Forma 3
Se iba	می¬رفت	میرفت	می رفت

En muchas ocasiones un espacio más en una palabra puede conducir a palabras con significados diferentes, según se puede observar en el ejemple de la Tabla 2.21.

Tabla 2.21: Espacio adicional cambiando el significado de una palabra

Palabra sin espacio	Traducción	Palabra con espacio	Traducción
مادر	madre	ما در	nosotros en
تنها	solo	تن ها	cuerpos

2.6.2.11. Confusión de las Letras en el Texto

Podemos escribir la letra " آ " (/a/, *alef* con *madda* arriba) al lugar de la letra " ا "(/a/, *alef* sin *madda*). Los ejemplos se pueden ver en la Tabla 2.22.

Tabla 2.22: Confusión de las letras

Palabra	Forma 1	Forma 2
Proceso	فرایند	فرآیند
África	افریقا	آفریقا
América	امریکا	آمریکا

Se puede utilizar o no la letra ‘ی’ en las palabras terminando en ‘ا’ (ver la Tabla 2.23).

Tabla 2.23: Letra ‘ی’ en lugar de ‘ا’

Palabra con "ی"	Palabra con "ا"	Pronunciación	Traducción
موسی	موسا	Moosa	Judío
عیسی	عیسا	Isa	Jesús
دکتری	دکترا	doctora	doctorado

Algunas letras persas se parecen mucho y en el caso de que una palabra tenga letras similares entonces, una letra puede cambiar el lugar que le corresponde por falta de cuidado. Por ejemplo, la palabra “زر” significa “oro” y tiene dos letras muy parecidas. Si cambiamos el lugar de cada letras solo poniendo el punto sobre la segunda letra, entonces tenemos la palabra “رز ” que significa “rosa”.

2.6.2.12. Diversidad de Ortografía o Escritura

Algunas palabras pueden ser escritas con letras diferentes y todas las formas son correctas como por ejemplo las que se muestran en la Tabla 2.24.

Tabla 2.24: Diversidad de ortografía

Palabra	Forma 1	Forma 2
Emperador	امپراتور	امپراطور
Habitación	اتاق	اطاق
Billete	بلیت	بلیط

Existen palabras que tienen sólo una forma correcta de escribir pero en algunos casos la forma incorrecta también se utiliza como por ejemplo los mostrados en la Tabla 2.25.

Tabla 2.25: Utilización de forma incorrecta de palabras

Palabra	Forma incorrecta	Forma correcta
Carbón	زغال	ذغال
Contento	خوشنود	خشنود

2.6.2.13. Ambigüedad Unicode

Hay algunas letras como "ى" [i,y] y "ك" [k] para los que tenemos más de un código Unicode (uno en persa y otro en árabe). Como algunas aplicaciones pueden utilizar diferentes Unicode entonces, tenemos que unificar sus ocurrencias antes del procesamiento del texto.

2.6.2.14. Ambigüedad de la Detección de los Nombres Propios

Puesto que no hay letras mayúsculas en la transcripción persa, la detección de los nombres propios en el texto acarrea algunos problemas. En la lengua persa no hay una regla general para distinguir los nombres propios de los otros sustantivos.

2.6.3. Lengua Persa y Procesamiento del Texto en la RI

Como se muestra en la Figura 1.2, la representación de documentos en un sistema de RI se define como algunas operaciones sobre el texto que dependen, sobre todo, de las características de lengua de documentos. Obtener una representación adecuada de un documento o consulta en un sistema de RI es una cuestión clave (Strzalkowski, et al., 1994). Por razones históricas, los documentos han sido generalmente representados como conjuntos de términos. Una de las operaciones sobre los términos consiste en identificar y eliminar algunos términos conocidos como palabras vacías que tienen un valor semántico muy escaso. La identificación de estas palabras para el texto persa se explica en detalle en el Capítulo 5. Otras operaciones como tokenización y lematización

tienen un aspecto sintáctico y presentan algunas dificultades en el texto persa que necesitan consideraciones especiales.

2.6.3.1. Segmentación del Texto

Segmentación del texto o *tokenización* es una de las primarias actividades en la construcción de un sistema de RI y se refiere al proceso de reconocimiento de los límites de los componentes del texto incluyendo oraciones, frases y palabras. La lengua persa tiene diferentes formas de escribir las palabras utilizando o eliminando el espacio banco y el uso de diversas formas de caracteres. Así, la tokenización y la conversión de estas formas y estilos en una norma única es un paso necesario en los sistemas del procesamiento del texto. Hay de una a cuatro formas de escribir una letra según su lugar en una palabra. El espacio no es un delimitador determinista y puede no ser un signo de límite para distinguir la palabra. Puede aparecer dentro de una palabra o entre palabras o puede estar ausente entre algunas palabras secuenciales. Por lo tanto, hay muchas palabras que se pueden escribir con el espacio, el espacio corto o ningún espacio.

Uno de los símbolos problemáticos en la lengua persa es el marcador "*Ezafe*". *Ezafe* es una vocal corta añadida entre preposiciones, sustantivos y adjetivos en una frase para determinar la relación entre los sustantivos y sus modificadores (sustantivos o adjetivos) o entre preposición y el sustantivo. Por lo general, *Ezafe* se pronuncia pero no está escrito en el texto y puede crear muchas ambigüedades sintácticas. *Ezafe* tiene una función parecida a la palabra "de" en español. En el caso de no poner *Ezafe* causa problemas en la identificación de las partes de la oración y frases cortas (como sintagmas nominales) y el procesamiento sintáctico y semántico del texto. Mientras que las diferentes formas de escribir *Ezafe* causa problemas de ambigüedades en la tokenización y la lematización del texto (Shamsfard, 2011).

Otro problema que ocurre en el proceso de tokenización del texto persa es la detección de los verbos. Ya se indicó anteriormente que los verbos simples son muy pocos en comparación con verbos compuestos (conocido como predicados complejos).

Estas construcciones consisten en un elemento pre-verbal (sustantivo, adjetivo o preposición) seguido por un verbo ligero tal como los verbos کردن [*kardan*] (hacer, en español), دادن [*dâdan*] (dar) o زدن [*zadan*] (pegar). En estas estructuras, el verbo pierde su significado original pero se asocia con otros elementos para formar un nuevo verbo. El significado de un verbo ligero es no composicional; es decir, no puede ser obtenido traduciendo cada elemento por separado. Por ejemplo, el verbo از دست دادن [*az dast dadan*] tiene tres partes separados con espacio. Si traducimos cada palabra tenemos "de mano dar" mientras que significa el verbo "perder" en español. Como podemos ver las tres partes deben estar juntas para tener un significado conjunto. Si las tres partes están separadas entonces cada palabra tiene su propio significado distinto. Desde el punto de vista semántico, el proceso de la tokenización debe juntar las diferentes partes entre sí para formar una sola palabra.

2.6.3.2. Lematización del Texto

El objetivo principal de la lematización es el de reducir las diferentes formas lingüísticas de una palabra a una forma común o *stem*, y así facilitar el acceso a la información durante el posterior proceso de búsqueda, paralelamente se está reduciendo el número de términos diferentes del sistema, lo que permite una reducción de los recursos de almacenamiento requeridos. El proceso de lematización en el texto persa tiene algunos problemas que son:

- Se utilizan cinco sufijos para formar palabras plurales y además hay palabras con plurales irregulares. Los sufijos pueden ser separados o adjuntos a la palabra. En el caso de que los sufijos estén agregados a la palabra, la distinción de los signos plurales es problemática, porque hay muchas palabras que se terminan con los signos plurales pero que no son términos plurales. En el caso de plurales irregulares, hay que buscar la forma singular en el léxico.
- Los sufijos que tienen papeles gramaticales de comparativos, pronombres personales y posesivos están generalmente agregados a las palabras. Hay muchas palabras que tienen estos sufijos pero que no pertenecen a estas categorías en una oración.

- En la lengua persa, las palabras se construyen generalmente a partir de la forma imperativa de los verbos. Por lo tanto, desde un punto de vista de la lingüística, la primera etapa para extraer la raíz es encontrar el modo imperativo de la palabra. En general, no es fácil obtener el modo imperativo ya que hay infinitivos irregulares. La forma imperativa del infinitivo irregular se basa en cómo se escuchan o se usan las palabras. En este caso, se necesita buscar el modo imperativo en el léxico.
- Por lo general, hay muchas excepciones gramaticales que deben ser consideradas en la implementación de los algoritmos de lematizadores.

2.7. Conclusiones

En un sistema de recuperación de información la parte de la representación de los documentos tiene estrecha relación con el procesamiento del texto de documentos almacenados. La complejidad asociada al lenguaje natural cobra especial relevancia cuando necesitamos recuperar información textual que satisfaga la necesidad de información de un usuario. Es por ello que en el área de RI textual son muy utilizadas las técnicas de procesamiento del lenguaje natural, tanto para facilitar la descripción del contenido de los documentos como para representar la consulta formulada por el usuario, y ello, con el objetivo de comparar ambas descripciones y presentar al usuario aquellos documentos que satisfagan en mayor grado su necesidad de información.

El idioma persa es uno de los idiomas con tareas complejas y desafiantes del procesamiento de texto. Las diferentes formas de la escritura son los principales problemas que necesitaran unificar o bien normalizar el texto para convertir todas las formas de escritura en una única norma estándar antes de procesar el texto. Esta tarea que la llamamos pre-normalización o preparación del texto es una tarea adicional antes del proceso de generación de los términos asociados a un documento en un sistema de RI.

La segmentación del texto y la definición del límite de las palabras son tareas muy difíciles. La mayoría de las letras tienen tres a cuatro formas de escritura. Cada

forma se utiliza en función de la posición de la carta dentro de la palabra que puede ser inicial, media o final (aislado). Hay varias escrituras para escribir textos persas que difieren en el estilo de escritura de palabras usando o eliminando los espacios dentro o entre las palabras utilizando diversas formas de caracteres. Entonces la correcta tokenizatión y la conversión de estas formas y estilos en una única norma es un paso necesario y adelante en la construcción de los sistemas de RI con documentos persas.

Una de las tareas en el proceso de la representación de los documentos en un sistema de RI es identificar y eliminar las palabras vacías. Por lo general, este conjunto de palabras se compone de preposiciones, artículos, adverbios y etcétera que aparecen en el texto con frecuencia, pero no llevan la información importante en términos de RI. En la estructura gramatical de la lengua persa cuando se trata del verbo compuesto, dichas palabras son elementos pre-verbales y se deben considerar junto al verbo para tener su significado. Por secuencia, estas palabras no tienen la función de palabras vacías y no deben ser eliminados.

La estructura de palabras persas requiere consideraciones especiales con el fin de encontrar sus raíces o lemas. La diversidad de formas plurales, plurales irregulares y las palabras no plurales terminando con los signos plurales son unos de los retos en la construcción de lematizadores. Los sufijos de comparativos, pronombres personales y posesivos están generalmente agregados a las palabras y hay muchas palabras que tienen estos sufijos pero que no pertenecen a estas categorías. La mayoría de los verbos son verbos compuestos para los que hay una dependencia de larga distancia. Es decir, los verbos compuestos están constituidos de diferentes partes ya separadas que se deben juntar para formar una palabra semánticamente correcta.

Parte II

Recuperación de Información en Persa

Capítulo 3

Estado del Arte de la Recuperación de Información en Persa

Resumen

En este capítulo presentamos un análisis de los últimos avances en la recuperación de información en relación con la lengua persa. En primer lugar, en la sección 3.2 revisamos todas las colecciones de documentos que fueron construidas para ser utilizadas en el ámbito de RI y procesamiento del texto persa. A continuación, describimos los temas que son dependientes del lenguaje de documentos en un sistema de RI. Estos temas son específicos a la lengua persa y necesitan investigaciones más profundas sobre sus características. La sección 3.3 presenta la construcción de tokenizadores del texto persa. En la sección 3.4 enumeramos las listas de palabras vacías identificadas por diferentes grupos de investigadores y la sección 3.5 analiza todos los algoritmos de lematización creados para la lengua persa. Por último, terminamos el capítulo con los métodos de indexación y modelos de recuperación de información que han sido aplicados a los documentos persas en la sección 3.6. Los temas descritos en esta última sección son de caracteres genéricos en el proceso de la RI y se pueden también aplicar a la lengua persa con pocas modificaciones.

3.1. Introducción

Este capítulo presenta el conjunto de conceptos, ideas e implementaciones relacionadas en el campo de RI en persa. Comentaremos todos los trabajos previos que son dependientes del lenguaje de los documentos en un sistema de RI textual. Aunque las principales técnicas propuestas en la RI dependen en mayor o menor medida del idioma de documentos y consultas que se formulan pero, denominamos la parte “más dependiente del lenguaje” a todas aquellas operaciones que se hacen desde la creación de un documento textual hasta la organización de la información (indexación). Todas estas operaciones se pueden resumir en el procesamiento del texto original que son: la

creación y almacenamiento de un documento, el análisis léxico (tokenización) del texto, la identificación de palabras vacías y la lematización.

A pesar de la naturaleza y las características especiales de la lengua persa, muy poco esfuerzo se ha dedicado a la RI en comparación con otros idiomas como el inglés. El campo de la RI es muy joven en Irán y hay pocos investigadores que trabajan en esta área. Hace poco tiempo que los investigadores han comenzado a tomar en consideración la investigación en el ámbito de la RI para los documentos persas. El uso de Internet ha crecido rápidamente en la sociedad iraní y resulta evidente que las herramientas y técnicas comercialmente disponibles para el idioma inglés no son tan fiables para la lengua persa y habrá que desarrollar herramientas específicas a la lengua.

Los modelos de la RI no son tan dependientes del lenguaje como la parte del procesamiento de texto. Estos modelos tienen un aspecto genérico y se pueden aplicar, de una manera general o con pocas modificaciones, a los documentos escritos con otros idiomas para recuperar informaciones. Es por eso que los modelos utilizados para los documentos persas son casi los mismos modelos que se han creados a lo largo del tiempo para otros idiomas como el inglés.

3.2. Colección de Prueba

A los efectos de evaluar los sistemas de RI completos o nuevos métodos y técnicas es necesario disponer de juegos de prueba normalizados (corpus con preguntas y respuestas predefinidas, corpus clasificados, etc.). Esta área tiene que ver con la producción de tales conjuntos, a partir de diferentes estrategias que permitan reducir la complejidad de la tarea, manejando la dificultad inherente a la carga de subjetividad existente. Hasta hace poco uno de los mayores problemas en el ámbito de la RI era la falta de colecciones de evaluación con suficiente entidad y de libre acceso, para que de este modo permitiesen una evaluación de los sistemas lo más completa posible y que dichos resultados fuesen comparables (Vilares Ferro, 2005).

En general, una colección de prueba tiene las tres siguientes componentes principales:

1. Un conjunto de documentos.
2. Un conjunto de pruebas de necesidades de información, expresadas como consultas efectuadas en lenguaje natural.
3. Un conjunto de juicios de relevancia para cada consulta, es decir, el conjunto de documentos que se consideran relevantes para todas las consultas contenidas en el segundo conjunto.

La elección de una colección adecuada es de la suma importancia a la hora de evaluar un sistema, ya que únicamente así se tendrá la convicción de que los resultados obtenidos son fiables y representativos. La calidad de una colección viene dada por diversos aspectos (Vilares Ferro, 2005):

- Su disponibilidad para la comunidad científica. El libre acceso a una colección promueve su utilización por otros investigadores, facilitando la comparación de resultados.
- El tamaño de la colección. Cuanto mayor sea el repositorio de documentos y el número de consultas a utilizar, más se ajustarán los resultados obtenidos al comportamiento real del sistema(Hull, 1996).
- La calidad de consultas. Dicha calidad depende de su variedad, de la diversidad de construcciones empleas, y de si dichas consultas se corresponden o no a necesidades de información realísticas.
- La calidad de los documentos. Viene dada por la variedad de los mismos y por su realismo en cuanto a que no hayan sido sometidos a ningún tipo de tratamiento especial.

Una colección de prueba es una herramienta experimental indispensable para los investigadores en RI ya que permite comprender la naturaleza de los resultados, compararlos con otros y reproducir pruebas en iguales condiciones. La mayoría de estas

colecciones se crean para el inglés u otros idiomas predominantes. TREC[8] (*Text REtrieval Conference*) es una de las colecciones de prueba bien conocida (Cacheda Seijo, et al., 2011). Dado que el proceso de la recuperación de información es altamente dependiente del lenguaje natural de los documentos, por lo tanto el método de evaluación así como las colecciones de prueba deben ser desarrollados para diferentes lenguas naturales. Con este fin, han sido creadas las colecciones CLEF[9] (*Cross-Language Evaluation Forum*) y *multi- lingual track* de *TREC*. Desafortunadamente, no hay información específica en persa en estas colecciones, a pesar de la considerable cantidad de contenido disponible actualmente en este idioma.

Experimentos con la lengua persa han sido bastante nuevos y limitados en comparación a los trabajos en otros idiomas. La mayoría de los investigadores de la recuperación de información y del procesamiento del lenguaje natural construyeron sus propias bases de datos que son por lo general pequeñas y recolectadas manualmente. La calidad de estos documentos no se ha estudiado por lo que no está claro cómo los resultados experimentales se pueden ampliar. Uno de los requisitos específicos para el procesamiento del texto persa es una colección estándar de prueba. Esta colección de prueba puede ser utilizada en diversas áreas de investigación como RI, *Text Mining*, procesamiento del lenguaje natural y otras. En lo que sigue en esta sección, describimos las colecciones de prueba o corpus existentes en la lengua persa.

3.2.1. Corpus Qavanin

Qavanin (las leyes, en español) es una de las primeras colecciones con documentos textuales en persa. Esta colección fue construida mediante la cooperación con la empresa DPI (*Data Processing of Iran*) el departamento del procesamiento de datos de Irán. La colección contiene las leyes y los reglamentos aprobados por el parlamento iraní durante 90 años. El tamaño de la colección es 25 MB y existe una gran variación

[8] http://trec.nist.gov/

[9] http://www.clef-campaign.org/home.html

en la longitud de los documentos. Por ejemplo, una simple ley con unos pocos párrafos se considera un documento y el presupuesto anual del gobierno conjunto con todas sus secciones y sub-secciones se identifican también como un documento único. Por lo tanto, en un trabajo realizado por Garamalek (Garamaleki, et al., 2002), los documentos se dividieron en fragmentos. Cada fragmento contiene una sección o sub-sección de una ley de unos pocos párrafos de longitud. De esta manera se ha construido una colección con 177.089 fragmentos.

Observaciones

El tamaño de esta colección es pequeño y hay sólo 15 consultas. Los documentos relevantes se descubren por un sistema de RI utilizando el método *pooling*. El enfoque *pooling* es analizar de manera manual un número determinado de documentos recuperados con distintos sistemas, este número suele ser elevado (varios centenares) y se corresponde con los primeros documentos recuperados con cada sistema. Este conjunto de documentos es el que de manera manual analizan los expertos, que son los encargados de decir en último término si son relevantes o no. Este sistema asume que la gran mayoría de los documentos relevantes son encontrados, si no por todos los sistemas, sí al menos por alguno de ellos, y los no recuperados pueden considerarse como no relevantes (Kowalski, 1997).

El juicio de relevancia de los documentos a una consulta en el corpus Qavanin se hace mediante la asignación de un valor entre cero y cuatro y una desventaja principal de esta colección es su dominio específico y su aplicación está muy limitada.

3.2.2. Colección de los Documentos de ISRI

Un conjunto de documentos textuales fue creado por ISRI (*Information Science Research Institute*) de la universidad de Nevada. La creación de esta colección era parte del proyecto "*Farsi Searching and Display Technologies*" elaborado por la universidad de Nevada (Taghva, et al., 2003a). Este corpus consta de 1.850 documentos de formato texto. Estos documentos fueron recogidos, dentro de un periodo de seis meses, de

algunos sitios web que publican noticias o revistas en persa. Algunos de estos sitios web están en Irán y presentan típicamente versión electrónica de los periódicos y revistas populares iraníes. El resto de los artículos son de sitios web que están en Estados Unidos o Europa. En estos documentos tratan temas relacionados con la política, económica, los problemas sociales y los cambios históricos en el oriente medio. Esta colección contiene 60 consultas y la relevancia del documento a cada consulta es binaria.

Observaciones

El propósito de crear esta colección era la comprobación de unas herramientas desarrolladas por ISRI y el corpus no está disponible en la Web para los investigadores. El tamaño de la colección es pequeño y los documentos no se han creados siguiendo las indicaciones de *TREC*.

3.2.3. Corpus Mahak

El corpus Mahak es otra colección de prueba en persa (Esmaili, et al., 2007). Este corpus contiene 3.007 documentos y 216 consultas. Para cada consulta fue construida una lista de los documentos relevantes aplicando diferentes métodos afín de mejorar su precisión,. Los documentos de este corpus son los artículos de noticias de ISNA (*Iranian Student's News Agency*). Entre las agencias en línea de noticias iraní, ISNA es la más antigua y se estableció en el año 2000. Aunque las noticias de ISNA están accesibles mediante su sitio web se proporcionan también en formato XML. En la Figura 3.1 podemos ver un documento de noticias de ISNA en formato de XML. Las características de Mahak son las siguientes:

- La distribución del tamaño de los documentos se extiende desde 44 hasta 52.086 bytes con 2,7 KB en promedio.
- La distribución del tamaño de las consultas tiene un promedio de 8,1 palabras por consulta con un mínimo de 2 y un máximo de 22 palabras en las consultas.

El corpus de Mahak se puede descargar desde http://ce.sharif.edu/~shesmail/Mahak/.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Root>
 - <News version="2.00" date="2006-06-03 03:51:46" lastupdate="2006-51-03 03:51:47" security="Normal" language="Persian"
    id="8503-07597" nid="729018" type="HyperText" pursuant_picture="Yes" pursuant_voice="No">
  - <Country>
     <Code>100</Code>
     <PName>ایران</PName>
     <EName>Iran</EName>
    </Country>
  - <City>
     <Code>100</Code>
     <PName>تهران</PName>
     <EName>Tehran</EName>
    </City>
  - <Subject>
     <Code>10</Code>
     <PName>سياسي</PName>
     <EName>Politic</EName>
    </Subject>
    <Title>حجت‌الاسلام‌والمسلمين سيدحسن خميني در جمع آذربايجاني‌ها: آذربايجان، برجدار و ميرات نهضت امام خميني (س) است</Title>
  - <Body>
     <![CDATA[ <P>ي از هموطنان آذري كه در قالب كاروان عاشورايي تبريز وارد تهران شده بودند، پس از حضور در حسينيه
     .هماران و تجديد ميثاق با آرمان‌هاي بنيانگذار جمهوري اسلامي، با حجت‌الاسلام‌والمسلمين سيدحسن خميني ديدار كردند
     <P>روان عزم، اراده، عشق و شور است و اين سبب شده كه بنده هر سال منتظر ديدن شما عزيزان در اين مكان مقدس باشم
     <P>و ارزشمندي است كه سبب شده پير، جوان و نوجوان گرد هم جمع شوند و به عشق امام (س) اين مسافت طولاني را طي كنند
     <P>.نوه امام تصريح كرد: مطمئن باشيد هر گامي كه شما در اين مسير مي‌پيمائيد هزاران ثواب در پي خواهد داشت
     <P>ي ايشان متعلق به همه ملت از هر قوم و نژادي است و همه بايد در حفظ و نگهداري انقلاب كه ميراث امام است، بكوشيم
     <P>يانجان، پرهدار انقلاب و نهضت امام خميني(س) مي‌باشد و اميدوارم مجون گذشته در حفظ وحدت و كيان اسلامي كوشا باشند
     <P>در ايام سالگرد ارتحال امام با پاي پياده از تبريز به ست تهران حركت، تا در مراسم سالگرد امام (س) شركت كنند
     <P>انتهاي پيام</P> ]]>
    </Body>
  </News>
</Root>
```

Figura 3.1: Noticias de ISNA en formato XML

Observaciones

La colección de Mahak fue creada por la universidad de Sharif[10] en virtud de investigaciones en la tarea de RI y no corresponde a las normas de TREC. La colección sólo tiene 3.000 documentos. Las etiquetas del documento XML no son muy representativas según TREC. Además, la relevancia de un documento a una consulta tiene un valor entre 0 y 100 mientras que el valor asignado a la relevancia, según las normas de TREC, debe ser el valor 0 o 1 (relevante o no relevante) sin ambigüedad.

[10] http://www.sharif.ir/web/en/

3.2.4. Corpus Hamshahri

Otra colección de documentos en persa es el corpus Hamshahri (AleAhmad, et al., 2009). Los documentos de la colección de Hamshahri son artículos de noticias del periódico Hamshahri que tratan varios temas. El periódico Hamshahri (ciudadano, en español) es uno de los primeros periódicos en línea en Irán. El periódico Hamshahri[11] comenzó a ofertar al público sus archivos en la Web desde el año 1996.

El corpus Hamshahri contiene artículos de noticias desde el año 1996 hasta 2003 y tiene más de 160.000 documentos, 65 consultas con sus juicios de relevancia y está públicamente disponible en la Web[12]. Los documentos están etiquetados por un número único de identificación, la categoría y la fecha de publicación. El tamaño de los documentos varía de noticias breves (menos de 1 KB) hasta más bien largos artículos (cerca de 140 KB) con una media de 1,8 KB. El tamaño global de la colección incluyendo etiquetas es 564 MB. Algunas características de esta colección se resumen en la Tabla 3.1.

Tabla 3.1: Características de la colección Hamshahri

Atributos	Valores
Tamaño de la colección	564 MB
Longitud de la colección	63.513.827 términos
Formato de documentos	texto
Número de documentos	166.774
Número de términos únicos	417.339
Longitud media de documentos	380 términos
Número de categorías	82
Número de temas	65

[11] Hamshahri newspaper, http://www.hamshahri.net/

[12] http://ece.ut.ac.ir/dbrg/hamshahri/

Todos los documentos están clasificados en 82 temas diferentes sobre la base de las categorías de noticias que los estaban disponibles en el sitio de Hamshahri. Cada documento de la colección está etiquetado con una palabra que indica su categoría. Como ejemplo, la etiqueta [siasi] que es la romanización de la palabra سیاسی en persa (política, en español). Aunque la colección de Hamshahri cuenta con 82 categorías, sólo 12 categorías contienen más de 1.000 documentos y cubren casi el 72% de la colección. Hay que mencionar que algunas categorías están muy relacionadas entre sí de modo que puede resultar conveniente fusionarlas y reducir el número de las categorías. Generalmente, las categorías representan temas en los cuales es necesario categorizar a los documentos facilitando su organización y posterior recuperación. El hecho de reducir el número de categorías facilita la clasificación de los documentos. La Tabla 3.2 muestra las 16 principales categorías de la colección con sus etiquetas cuyas categorías contienen más documentos que otras categorías.

Tabla 3.2: Categorías principales en la colección Hamshahri

Etiqueta de categorías	Nombre de categoría en persa	Nombre de categoría en español
adabh	ادبی-هنری	Literatura-arte
akhar	اخبارکوتاه	Breves noticias
bankb	بورس و بانک	Mercado y banco
econw	اقتصاد جهانی	Economía mundial
ejtem	اجتماغی	Sociedad
elmif	علمی و فرهنگی	Ciencia y cultura
eqtes	اقتصادی	Economía(en Irán)
gards	گردش گری	Turismo
gungn	گوناگون	Diversos
havad	حوادث	Eventos sociales
ikaba	فناوری اطلاعات	Tecnología de información
kharj	اخبار خارجی	Noticias de exteriores
shahr	شهر تهران	Teherán y asuntos municipales
shrst	شهرستانها	Ciudades en Irán
siasi	سیاسی	Política
Vrzsh	ورزشی	Deporte

La Figura 3.2 muestra un tema y la Figura 3.3 indica un documento del corpus Hamshahri en formato XML.

```
<topic lang="fa">
<identifier>10.2452/551-AH</identifier>
<title>تنیس جام ویمبلدون</title>
<description>نام برندگان جام</description>
<narrative>فاتحان مرد و زن جام ویمبلدون چه کساني بوده اند</narrative>
</topic>
```

Figura 3.2: Ejemplo de un tema de la colección Hamshahri

```
<DOC>
<DOCID>H-750402-15S1</DOCID>
<DOCNO>H-750402-15S1</DOCNO>
<DATE>1996-06-22</DATE>
<CAT xml:lang="fa">اقتصاد</CAT>
<CAT xml:lang="en">Economy</CAT>
<TEXT>
كاهش مجدد مازاد بازرگاني ژاپن
واحد رسانه هاي خارجي همشهري: مازاد تراز تجاري ژاپن براي هجدهمين
ماه متوالي در ماه مه ۱۹۹٦ كاهش يافت.
به گزارش تلويزيون سي. طبق ان ان، اظهار گمرك ژاپن، مازاد تراز
تجاري اين كشور در اين ماه نسبت به ماه مشابه سال /٦٠ ٥قبل درصد
كاهش نشان مي دهد.
ميزان رشد واردات ژاپن بسياربيشتر از صادرات بود و درنتيجه
مازاد تجاري اين كشور به /٢ ١٢ميليارد دلار رسيد. همچنين مازاد
تجاري دوجانبه با آمريكا هم، كه از نظر سياسي بسيار حساس است، قريب
به ٤٠ درصد پايين آمد.
</TEXT>
</DOC>
```

Figura 3.3 Ejemplo de un documento en formato XML en la colección Hamsahhri

La versión original de Hamshahri fue creada en el año 2007. Sus documentos están en textos planos y almacenados en formato de *TREC*. Otra versión (versión 1 de CLEF2008) fue preparada en el año 2008, conteniendo 100 temas con juicio de

relevancia y el mismo número de documentos que la versión original. La segunda versión (versión 2 de CLEF2009) fue preparada en el año 2009 y contiene alrededor de 320.000 documentos y 50 temas con sus juicios de relevancia. Los documentos de esta última versión también contienen imágenes. La Tabla 3.3 resume las especificaciones de dos últimas versiones de Hamshahri.

Observaciones

Las dos últimas versiones de Hamshahri están construidas según las especificaciones de *TREC* y tienen alrededor de 320.000 documentos y 150 temas en formato XML con juicio de relevancia. El corpus Hamshahri es una de las mejores fuentes de documentos en persa para los investigadores en el ámbito de RI.

Tabla 3.3: Características de dos últimas versiones de Hamshahri

Criterios	Versión 1	Versión 2
Tamaño (Unicode CLEF formato XML)	700 MB	1400 MB
Número de documentos	160.000 +	318.000 +
Intervalo de tiempo de los documentos	del 23.04.1996 al 11.02.2003	del 23.04.1996 al 13.05.2007
Categoría de documentos	sí	sí
Enlace a imágenes	no	sí
Enlace a las páginas web originales	no	sí
Consultas + juicio de relevancia	sí	sí
Número de consultas	100	50

3.2.5. Corpus Bijankhan

Bijankhan es un corpus etiquetado orientado a la investigación en el ámbito de procesamiento del lenguaje natural en la lengua persa (Oroumchian, et al., 2006). Esta colección se ha construida con noticias y textos comunes. Todos los documentos se clasifican en diferentes temas como la política, cultura, etc. y hay en total 4.300 diferentes temas. La colección Bijankhan contiene aproximadamente 2,6 millones de palabras manualmente etiquetadas con un conjunto de etiquetas que contiene 550

etiquetas de POS (*Part-Of-Speech*). *Part of speech tagging* consiste en anotar cada palabra de un texto con la mayoría categoría sintáctica adecuada.

POS tagging tiene dos pasos principales. El primer paso es encontrar el posible conjunto de etiquetas de cada palabra, independientemente de su función en la oración y el segundo paso es la elección de la mejor etiqueta entre las etiquetas posibles en función de su contexto. Esta colección fue preparada y distribuida por el grupo de investigación de base de datos en la universidad de Teherán. En agradecimiento al profesor Bijankhan de la facultad de literatura y ciencias humanas de la universidad de Teherán, se eligió su nombre para esta colección por sus aportaciones en la versión original del corpus. El corpus Bijankhan se puede descargar desde la página web, http://ece.ut.ac.ir/dbrg/Bijankhan/.

Observaciones

Dada la estructura de la colección de Bijankhan, este corpus es un corpus lingüístico y más apropiado para la investigación en el dominio de procesamiento del lenguaje natural que en el dominio de la RI.

3.2.6. Resumen

La Tabla 3.4 indica una breve comparación entre las diferentes colecciones de prueba que existen para la lengua persa.

Tabla 3.4: Comparación de diferentes colecciones de documentos persas

Corpus	N° de documentos	Tamaño (MB)	N° de temas	Disponibilidad
Qavanin	177.089	25	15	no
ISRI	1.850	-	60	no
Mahak	3.007	-	216	sí
Hamshahri (versión original)	166.774 en texto	564	65	sí
Hamshahri (versión 1 CLEF2008)	166.774 en XML	700	100	sí
Hamshahri (versión 2 CLEF2009)	318.000 en XML	1400	50	sí
Bijankhan	-	149	-	sí

Qavanin es el conjunto de las leyes y reglamentos aprobados por el parlamento iraní. Esta colección de documentos es de dominio específico y su aplicación está muy limitada. Bijankhan es un corpus lingüístico que es más apropiado para la investigación en el dominio de procesamiento del lenguaje natural en la lengua persa. Las colecciones de ISRI y Mahak son muy pequeñas y no están construidas según las especificaciones de TREC. Entre todas las colecciones de prueba en persa, el corpus Hamshahri es una colección estándar que fue construida en 2007. Dos otras versiones (versión 1 de CLEF2008 y versión 2 de CLEF2009) fueron creadas en 2008 y 2009 y desde entonces se utilizan como una fuente de documentos persa para los investigadores de RI. La colección de Hamshahri fue construida según las especificaciones de TREC y su tamaño es relativamente más grande que otras colecciones. Las dos últimas versiones tienen alrededor de 320.000 documentos y 150 temas en formato XML con juicio de relevancia.

3.3. Tokenización del Texto Persa

Antes de hacer cualquier operación sobre los documentos en un sistema de RI textual, hay que determinar el límite de palabras en el texto. Entonces, el proceso de análisis léxico o *tokenización* es el primero paso a realizar en el procesamiento del texto. Esta operación consiste en la conversión de una secuencia de caracteres, el texto de los documentos o consultas, en una secuencia de palabras o *token*s, ya que las palabras y frases identificadas en esta fase constituirán las unidades fundamentales sobre las que deben trabajar las etapas posteriores en un sistema de RI. En lenguajes artificiales la definición de lo que puede ser considerado un *token* puede ser precisa y definida sin ambigüedad. En cambio, los lenguajes naturales muestran una amplia variedad y muchos caminos para decidir sobre lo que puede ser considerado una unidad computacional para llegar a un texto.

Un *token* es una palabra del texto, la dificultad radica en cómo se pueden separar las palabras correctamente. Esta dificultad depende, en primer lugar, del idioma del texto que presentará una serie de fenómenos lingüísticos particulares a considerar.

Habría que resolver claramente qué es una palabra. En la mayoría de los lenguajes naturales un espacio en blanco o los signos de puntuación son un límite de la palabra pero no es el caso de la lengua persa. La lengua persa debido a sus morfologías complejas, diferentes formas de letras, uso o eliminación de los espacios y el uso de diversas formas de caracteres en una palabra, presenta muchas dificultades en la identificación de las palabras en un texto. No se podría decir que una palabra es una cadena de caracteres con un espacio antes y después, sin duda el tratamiento es un poco más complejo. Por ejemplo, si consideramos el texto siguiente (ver la Figura 3.4) en un software del proceso de textos como *Microsoft Word*, el número total de las palabras contadas es 67 palabras pero, en realidad, las palabras significadas son 53.

وب جهانی در ابتدا رسانه ای محسوب می شد که چیزی بیش از متن در خود نداشت.
وب سایت های امروزی می توانند شامل قابلیت های بسیاری از جمله تصاویر گرافیکی، صوت، انیمیشن، ویدیو و سایر مطالب چند رسانه ای می باشند.
زبان های اسکریپت نویسی وب مانند جاوا اسکریپت یکی ازساده ترین روشهای ایجاد رابطه متقابل با کاربران و خلق جلوه های دینامیکی محسوب می شود.

Figura 3.4: Un texto persa

Las palabras subrayadas están separadas por un espacio blanco pero forman una única palabra. Como ya indicamos (ver el Capítulo 2) a propósito de las propiedades de la lengua persa, los tokenizadores construidos para las lenguas tal como inglés no pueden ser fiables para el texto persa. La lengua persa necesita su propio tokenizador que considere las características específicas de la lengua para separar correctamente las palabras del texto. La correcta segmentación y separación de las palabras son muy importantes en un sistema de RI debido a que estas palabras serán candidatas a ser adoptadas como términos de índice por el sistema.

Esta sección se dedica a describir los esfuerzos y trabajos que se han producidos sobra la tokenización del texto persa.

3.3.1. Tokenización en el Proyecto Shiraz

El primer trabajo de tokenización fue un tokenizador desarrollado por Megerdoomian (Megerdoomian, et al., 2000). La construcción de un tokenizador era parte del proyecto Shiraz[13]. El tokenizador tiene dos sub-tokenizadores. El primero es un tokenizador de bajo nivel que es independiente del lenguaje para separar los elementos textuales en *tokens* básicos. La separación de las palabras se hace según los marcadores de puntuación o el espacio en blanco. El segundo, post-tokenizador, tiene informaciones especificas del idioma y se aplica luego sobre las palabras de salida del tokenizador de bajo nivel. Los algoritmos de post-tokenizador se utilizan principalmente para unir elementos de inflexión que han sido separados por el tokenizador al bajo nivel. Por ejemplo, como hemos indicado en la parte de la definición de la lengua persa considerando la forma final de un carácter entonces, una palabra se puede ser separada por diferentes partes. En esta situación, las diferentes partes se deben considerar juntos por el post-tokenizador para formar una correcta palabra. El algoritmo de post-tokenizador utiliza las reglas gramaticales y los datos específicos de la lengua en forma de tablas internas.

Observaciones

El hecho de que el tokenizador no es disponible en la Web hace que no se pueda aplicar al texto para hacer una evaluación de los algoritmos construidos. Constatamos que la lista de prefijos y sufijos separables que el post-tokenizador utiliza para volver a colocar los morfemas separados no es completa. Otro punto a tener en cuenta es que el post-tokenizador considera sólo dos *tokens* consecutivos y después cada token se va a comprobar con la lista de morfemas para poder unir los dos *tokens*. Mientras que en el texto persa hay, a veces, más de dos *tokens* consecutivos que son dependientes entre ellos y se deben considerar todos juntos. Además, se deben incorporar en el tokenizador

[13] El objetivo del proyecto de traducción automática Shiraz (http://crl.nmsu.edu/shiraz) fue la construcción de un prototipo de sistema que traduce texto persa al inglés. El proyecto se inició en octubre de 1997 y la versión final fue entregado el agosto de 1999. Este trabajo fue elaborado por el laboratorio de investigación en computación de la universidad del estado de nuevo México.

las especificaciones para desambiguar los límites de la frase y el reconocimiento de las siglas y abreviaturas.

3.3.2. STeP-1

STeP-1 es un conjunto de herramientas en el ámbito de procesamiento del lenguaje natural elaborado por el laboratorio de investigaciones de procesamiento del lenguaje natural de la universidad de Shahid Beheshti de Teherán (Shamsfard, et al., 2010). Uno de sus componentes tiene la función de hacer la tokenización y la correcta segmentación del texto. El enfoque propuesto combina métodos basados en diccionarios y reglas gramaticales y convierte diversas formas prescritas de escritura a una forma única estándar. El tokenizador desarrollado determina los límites de las palabras, reconoce los verbos de múltiples partes, números, fechas, abreviaturas y algunos nombres propios. El tokenizador utiliza una base de datos que contiene 57.000 elementos incluyendo nombres, adjetivos, verbos, prefijos y sufijos organizados en diferentes tablas. El tokenizador tiene las siguientes etapas:

1. Unificar la codificación de caracteres. Algunas letras persas pueden codificarse en Unicode persa o árabe. En tales casos, en el primer paso se convierte todo el Unicode árabe a su Unicode equivalente en persa.
2. Dividir el texto de entrada debido a la posición de los espacios en blanco y los signos de puntuación.
3. Separar los números (enteros o flotantes), fechas y letras inglesas del texto por un espacio.
4. Ajustar los espacios alrededor de los signos de puntuación.
5. Reconocer los verbos, sustantivos y adjetivos. En este paso se comprueba el *token* en el léxico. Si no está presente entonces se aplica algunos métodos de lematización para eliminar los afijos y se verifica el lema en el léxico. Los sufijos de verbos, signos plurales de los sustantivos y marcadores comparativos de los adjetivos son algunos de los afijos que deben ser eliminado.

6. Reconocer abreviatura. Letras individuales en persa y en inglés se reconocen como abreviaturas.
7. Reconocer números y palabras en inglés.
8. Procesamiento de las palabras indefinidas. Puede haber todavía unas palabras que no han sido reconocidas en el módulo anterior. En este caso, hay dos soluciones alternativas, lematización o insertar espacios blancos. El tokenizador sólo elimina algunos afijos específicos limitados. Para las palabras más complejas el sistema utiliza un lematizador. El lematizador no solo elimina los afijos pero también considera los cambios ortográficos durante la inflexión o derivación. La inserción del espacio blanco se efectúa cuando algunas cadenas desconocidas pueden ser convertidas a unas palabras conocidas mediante la inserción de espacios en lugares apropiados justo en el límite de las palabras.
9. Reescribir las palabras con diferentes ortografías o estilos en una única escritura estándar.
10. Convertir los espacios entre las partes de una palabra en un espacio corto (o nada) para concatenar partes separadas de una sola palabra.
11. Reconocer los verbos de múltiples partes. Cuando las diferentes partes de un verbo se deben considerar juntos.
12. Manejar las ambigüedades. Se trata de clarificar las ambigüedades con la partícula verbal imperfectiva y el marcador indefinido.

El componente de tokenización de STeP-1 fue comprobado, en primero, sobre 400 palabras. El rendimiento del tokenizador para distinguir correctamente las palabras fue alrededor de 87%. La segunda prueba fue aplicado sobe un texto con 100 frases. El texto contenía 80 afijos derivados, 61 prefijos verbales, 15 sufijos verbales, 10 acrónimos y abreviaturas, 34 fechas y números y 16 palabras concatenadas. El rendimiento del sistema fue de 86,6% de éxito.

Observaciones

El lematizador utiliza una base de datos para hacer la correcta segmentación del texto. La construcción de la base de datos es un trabajo enorme sobre todo si queremos contener todo el léxico persa. A pesar de ello, el rendimiento no es muy alto. Debido a

que el algoritmo de lematización utiliza una base de datos, el rendimiento debería ser mayor y eso se puede explicar que hay errores en la segmentación de palabras desconocidas para encontrar sus partes. Otros errores pueden ser debido a la falta de léxico y la dependencia de larga distancia entre las partes de los verbos compuestos. Es decir, en el caso de los verbos compuestos, hay muchas palabras consecutivas que son dependientes y se deben considerar junto para formar una sola palabra. A demás, la evaluación sobre solo dos textos no puede ser muy representativa.

3.3.3. Resumen

La Tabla 3.5 se resume los dos tokenizadores para el texto persa. Ambos algoritmos deben mejorarse para poder hacer la correcta segmentación del texto. El tokenizador de STeP-1 tiene más funcionalidad pero el proceso de tokenización es lento por el hecho de utilizar una base de datos.

Tabla 3.5: Tokenizadores del texto persa

Tokenizador	Uso de base de atos	Identificación de número, hora y fecha	Identificación de siglas y abreviaturas
Proyecto Shiraz	no	no	no
STeP-1	sí	sí	sí

3.4. Palabras Vacías en Persa

En la RI, tradicionalmente un documento se indexa y se busca por palabras (Baeza-Yates, et al., 1999). Por definición, las palabras vacías (*stopwords*, en inglés) son palabras muy comunes que aparecen en el texto con frecuencia, pero no llevan la información importante en términos de RI. Este conjunto de palabras se compone de preposiciones, artículos, adverbios, conjunciones, posesivos, demostrativos, pronombres, algunos verbos (por ejemplo estar y ser, en español) y algunos nombres. Las palabras vacías pueden afectar a la eficacia de recuperación debido a una muy alta frecuencia y tienden a disminuir el impacto de las diferencias de frecuencia entre las

menos palabras comunes. También pueden afectar a la eficiencia, lo que resulta en una gran cantidad de procesamiento improductivo. Estas palabras supondrían típicamente una parte importante del índice si no se quitaran en el momento de la indexación (Frakes, et al., 1992), lo que permite reducir considerablemente el espacio de almacenamiento de las estructuras generadas. En experimentos citados por Witten (Witten, et al., 1999) dicho ahorro supuso en torno a un 25%, mientras en (Baeza-Yates, et al., 1999) se habla de un 40% o incluso más. Según Kucera (Kucera, et al., 1982), las diez palabras más frecuentes en inglés representan entre el 20 y el 30% de los términos en un documento.

Por lo tanto, la identificación de una lista de palabras vacías que contiene dichas palabras para eliminarlas del procesamiento de texto es esencial para un sistema de la RI. Lista de palabras vacías se puede dividir en dos categorías; independiente de dominio y dependiente de dominio. Esta lista puede ser creada utilizando clases sintácticas o mediante la estadística de corpus, que es un enfoque más dependiente de dominio. También se puede crear usando una combinación de las clases sintácticas y estadísticas del corpus para obtener los beneficios de ambos enfoques (El-Khair, 2006).

En esta sección se describe todos los trabajos previos que llevaron a identificar las palabras vacías para el texto persa.

3.4.1. Palabras Vacías por el ISRI

Como parte de una serie de investigaciones elaboradas por el instituto de investigación en ciencias de información en la universidad de Nevada, hay un informe técnico que presenta una lista de palabras vacías en persa (Taghva, et al., 2003b). Para poder identificar las palabras vacías, en primero, se construyó una colección de 1.850 documentos que ya lo hemos citado en el parte de corpus (ver la Sección 3.2.2). Basando en la distribución de los términos en la colección, fue identificada una lista de palabras con altas frecuencias. Después, refiriendo a la lista de palabras vacías de inglés y al sentido común de las palabras, fue editado manualmente el resultado para eliminar

palabras que, aunque frecuente en la elección, no deben considerarse como palabras vacías en una colección general. Entre las palabras vacías identificadas hay también 12 verbos considerados como palabras vacías verbales que están mostrados en la Tabla 3.6. La lista final contiene 155 palabras vacías que está presentada en la Tabla 3.7. Para facilitar la comprensión a los lectores, la traducción de palabras vacías se muestra en la Tabla 3.8.

Tabla 3.6: Palabras vacías verbales identificadas por ISRI

Traducción en español	Verbos
hacer	کردن
ser, estar	بودن
hacerse, llegar a ser	شدن
tener	داشتن
querer	خواستن
decir	گفتن
dar	دادن
tomar	گرفتن
venir	آمدن
poder	توانستن
encontrar, buscar	یافتن
llevar	آوردن

Observaciones

La lista de palabras vacías de ISRI se deriva de una colección muy pequeña y como se ha mencionado por los autores mismos, es relativamente corta e incompleta. Algunas palabras fueron manualmente seleccionadas tomando como referencia la lista de palabras vacías de inglés y considerando la semántica de la palabra. Por ejemplo la palabra "جناح" [jenah] (lateral, en español) que está incluida en esa lista, se utiliza poco en el texto persa pero se puede considerar semánticamente como una palabra vacía.

Tabla 3.7: Palabras vacías identificadas por ISRI

دیگران	همچنان	مدت	چیز	سایر
جا	طی	کل	کنونی	بیرون
مثلا	کامل	کاملا	آنکه	موارد
واقعی	امور	اکنون	بطور	بخشی
تحت	چگونه	عدم	نوعی	حاضر
وضع	مقابل	کنار	خویش	نگاه
درون	زمانی	بنابراین	تو	خیلی
بزرگ	خودش	جز	اینجا	مختلف
توسط	نوع	همچنین	آنجا	قبل
جناح	اینها	طور	شاید	ایشان
جهت	طریق	مانند	پیدا	ممکن
کسانی	جای	کسی	غیر	بی
قابل	درباره	جدید	وقتی	اخیر
چرا	بیش	روی	طرف	جریان
زیر	آنچه	البته	فقط	چیزی
چون	برابر	هنوز	بخش	زمینه
بین	بدون	استفاده	همان	نشان
بسیاری	بعد	عمل	روز	اعلام
چند	آنان	بلکه	امروز	تمام
بیشتر	آیا	برخی	علیه	دیگری
ویژه	گذشته	انجام	حتی	داده
راه	سوی	ولی	زمان	حال
تنها	بسیار	یعنی	عنوان	همین
هیچ	پیش	وی	یکی	اینکه
وجود	شما	پس	چنین	میان
مورد	چه	اگر	همه	نه
دیگر	آنها	باید	هر	او
ما	من	تا	نیز	اما
یك	خود	بر	یا	هم
را	این	با	آن	برای
و	در	به	که	از

Tabla 3.8: Traducción en español de las palabras vacías identificadas por ISRI

Palabra	Palabra	Palabra	Palabra	Palabra	Palabra
otro	cosa	momento	todavía	otros, otras	ambos
fuera	actual	todo, toda	durante	sitio	para
casos	quien	perfectamente	perfecto	ejemplo	desde
parte, parcial	así que	ahora	asunto	verdad	o
presente	tipo	nada	cómo	abajo	aquello, aquella
mirada	si mismo	junto a	opuesto	posición	que
muy	tu	entonces	tiempo	interior	en
diferente	aquí	sin	si mismo	grande	con
antes	ahí	también	tipo	vía	en, a
ellos, ellas	tal vez	manera, modo	estos	lateral	si mismo
posible	claro	como	así	sentido	esto
sin	excepto	alguien	lugar	cualquiera	en
reciente	cuando	nuevo	acerca de	capaz	uno, una
pasando	lado	sobre	más	por qué	ra (marcado del objeto)
alguno, alguna	sólo	por supuesto	qué	bajo	y, e
base	parte	todavía	igual	debido a	pero
mostrar	mismo	uso	sin que	entre	ambién
declarar	día	acto	luego	muchos	hasta
completo	hoy	más bien	aquellos, aquellas	varios	yo
otro, otra	contra	unos, unas	si es	más	nosotros
dado	aunque	hecho	pasado	especial	el, ella
ahora	tiempo	pero	hacia	camino	cada
también	titulo	es decir	mucho(s)	sólo	debe
de esto, de esta	uno	el, ella	antes que	ninguno	esos
entre	tal	entonces	usted, vosotros	hay	otro
no	todo, toda	si	qué	caso	-

3.4.2. Palabras Vacías del Corpus Hamshahri

Además del corpus Hamshahri, los autores presentaron también algunas propiedades estadísticas de la lengua persa (AleAhmad, et al., 2009). Una de las propiedades fue la identificación de las letras y palabras con alta frecuencia en el corpus. Entre las letras identificadas sólo hay una letra que tiene significado en la lengua persa y es la letra و [va] (y, en español).

Tabla 3.9: Palabras de altas frecuencias en el corpus Hamshahri

Palabra	Frecuencia	Traducción	Palabra	Frecuencia	Traducción
و	2.821.897	y, e	برای	395.355	para, por
در	2.141.883	en	کشور	205.251	país
به	1.878.546	a	کنند	137.895	hacer
از	1.454.614	desde	دارد	133.926	tiene
که	1.173.433	que	کرده	124.457	hecho
می	1.066.084	partícula verbal de imperfectivo	باید	119.748	debe
را	892.063	ra (marcador del objeto)	قرار	115.268	acuerdo
با	740.039	con	مورد	110.859	caso
آن	329.195	eso, esa	آنها	109.628	aquellos
یک	306.437	uno	باشد	107.346	ser, estar
ها	258.762	signo plural	ایران	196.958	Irán
این	1.061.288	este, esto, esta	تهران	132,024	Teherán
است	914.627	es	خواهد	111.861	querer
های	573.600	signo plural + *ezafe*	عنوان	89.285	titulo
شود	284.184	ser, estar (derivado)	گزارش	83.671	informe
شده	275.384	ser, estar (derivado)	اعلام	76.267	declara
خود	265.807	si mismo	سیاسی	71.572	política
کرد	245.644	hecho	انجام	69.139	hecho
سال	203.846	año	هستند	68.524	son
کند	188.752	hacer (derivado)	گذشته	68.219	pasado
گفت	183.903	dijo	-	-	-

La Tabla 3.9 muestra los términos de alta frecuencia en el corpus Hamshahri con sus traducciones en español y está ordenada de acuerdo con el número de letras que se encuentra en la palabra.

Observaciones

Entre las 50 letras y palabras de alta frecuencia en el corpus, hay 41 palabras significativas. Las palabras, por ejemplo como سیاسی [siasi] (política, en español), a pesar de sus altas frecuencias en el corpus no se pueden considerar como palabras vacías. Estas palabras son dependientes del dominio y por lo general no son palabras vacías. En el conjunto de esta colección existe también un archivo con más de 800 palabras. Este archivo contiene los términos más frecuentes en la colección que también incluye los signos de puntuación y otros términos que no pueden ser considerados como palabras vacías.

3.4.3. Palabras Vacías del Corpus Mahak

Otro trabajo similar es la identificación de palabras con alta frecuencia en la colección Mahak (Esmaili, et al., 2007). Esta lista sólo contiene 35 palabras que, dado sus significados, pueden ser consideradas como palabras vacías. En la Tabla 3.10 presentamos todas las palabras y ponemos la traducción en español para facilitar la comprensión a los lectores.

Observaciones

La identificación de palabras vacías del corpus Mahak está basada en la alta frecuencia de palabras en la colección. El corpus es muy pequeño (3.007 documentos) y la lista obtenida es muy corta e incompleta para poder ser utilizado de forma general.

Tabla 3.10: Traducción de las palabras de alta frecuencia en el corpus Mahak

Palabra	Traducción	Palabra	Traducción	Palabra	Traducción
و	y, e	ما	nosotros	نیز	también
در	en	هم	ambos	به	en
به	a	هر	cada	سال	año
از	desde	یا	o	بود	fue
که	que	دو	dos	کار	trabajo
با	con	این	eso, este	شود	ser, estar
را	ra (marcador del objeto)	است	es	اما	pero
آن	eso, esa	کرد	Hacer(derivado)	داد	dar
یک	uno	گفت	dijo	روز	día
بر	en	وی	el, ella	نظر	vista
شد	ser, estar (derivado)	خود	si mismo	اگر	si
تا	hasta	شده	ser, estar	-	-

3.4.4. La Lista de Palabras Vacías por Davarpanah

La última lista de palabras vacías fue creada por Davarpanah (Davarpanah, et al., 2009). Esta lista es una agregación de dos listas de palabras vacías, una basada en el dominio y la otra en el corpus. En primero, un conjunto de 248.552 palabras fue creado por el procesamiento de 63 artículos textuales extraídos de 12 diferentes revistas de piscología, educación, biblioteca y ciencias de la información. Después, una lista de 1.291 palabras fue construida considerando la función sintáctica y frecuencia de cada palabra. De acuerdo con la recomendación de expertos, los términos potenciales de indexación o búsqueda fueron eliminados y se quedó una lista de 746 palabras. En la siguiente etapa, fue construida una lista de 758 palabras de alta frecuencia (más de 20.000 ocurrencias) utilizando el corpus Hamshahri. Por lo tanto, una comprobación manual se hizo para eliminar las palabras que no pueden ser consideradas como palabras vacías convencionales. La lista obtenida como la segunda lista ha tenido 422 palabras. Finalmente, las dos listas se agregaron juntas para generar la lista definitiva. Después de eliminar 246 palabras comunes entre las dos listas se quedan 922 palabras incluyendo también las palabras vacías verbales.

Observaciones

Los elementos de la primera lista fueron identificados de una colección muy pequeña de sólo 63 artículos textuales. El criterio de la selección de la segunda lista era la alta frecuencia de más de 20.000 sin justificar este límite en el número de ocurrencias. Es evidente que la elaboración de esta lista es manual y los expertos combinan las palabras vacías ya identificadas para extender la lista de palabras vacías. Por ejemplo la palabra از این که [az in ke] (cual o cuyo, en español) se encuentra en la lista definitiva que es la combinación de tres palabras vacías ya incluidas en la lista final. Estas palabras son از [az] (desde) + این [in] (este) + که [ke] (que).

3.4.5. Resumen

La Tabla 3.11 resume lo conjunto de palabras vacías persas ya identificadas por diferentes grupos de investigación.

Tabla 3.11: Lista de palabras vacías en persa

Lista de palabras vacías	N°. de palabras vacías (verbal y no verbal)	Método de identificación
ISRI	155+12(verbos)	alta frecuencia en la colección + manual
Hamshahri	41	alta frecuencia en el corpus
Mahak	35	alta frecuencia en el corpus
Davarpanah	922	alta frecuencia en el corpus+ semántico + manual

La identificación de listas de palabras vacías persas y disponibles se basa en la alta frecuencia de palabras en unas colecciones de textos. Hay algunas que son muy cortas e incompletas y otras que fueron construidas manualmente considerando el significado de la palabra. Sin embargo, el uso de una única lista de palabras vacías a través de diferentes colecciones de documentos persas podría ser perjudicial para la eficacia de recuperación en un sistema de RI. Por le tanto, es preferible poder derivar una lista de palabras vacías para una determinada colección. Esta es la razón por la que

se propone en el capítulo 5 un método automático para identificar las palabras vacías para sistemas de RI en persa.

3.5. Algoritmos de Lematización para la Lengua Persa

Una característica del lenguaje humano es la de que un mismo concepto puede ser formulado de maneras diferentes, que denominaremos variantes (Jacquemin, 1999). Esto supone que a la hora de la comparación de documentos y consultas nos podemos encontrar con que aún refiriéndose a conceptos equivalentes, al menos desde el punto de vista de un sistema de RI, puedan no producirse correspondencias debido a que ambos estén empleando términos diferentes (Hull, 1996). Para minimizar en lo posible el impacto de estos fenómenos, los sistemas de RI recurren a técnicas de lematización (*stemming*, en inglés) implementadas mediante herramientas denominadas lematizador (*stemmers*, en inglés).

La lematización consiste en la obtención de la raíz de las palabras, de forma que el proceso de indexación se lleve a cabo sobre ellas en lugar de sobre las palabras originales. Asumiendo que dos palabras que tengan la misma raíz representan el mismo concepto, esta técnica permite a un sistema de recuperación de información relacionar términos presentes en la consulta y en los documentos que pueden aparecer bajo diferentes variantes morfológicas. Si bien el objetivo principal de la lematización es el de reducir las diferentes formas lingüísticas de una palabra a una forma común o lema, y así facilitar el acceso a la información durante el posterior proceso de búsqueda, paralelamente se está reduciendo el número de términos de indexación es decir el espacio de almacenamiento, aumentar las tasa de recuperación (*recall*, en inglés) y la velocidad del proceso de búsqueda.

En el campo de la RI se han realizado muchos experimentos para determinar el valor de la lematización en el proceso de la recuperación. Hay una variedad de los métodos para construir un lematizador. El algoritmo extensamente usado de Porter (Porter, 1980) es un sistema basado en las reglas, que quita de una manera iterativa los sufijos. El algoritmo del Porter no garantiza la forma correcta de las palabras que se

producirán después de lematizacion. Sin embargo, su algoritmo es consistente y se demuestra que aumenta la recuperación hasta 15%. Kraaij (Kraaij, et al., 1995) ha demostrado que el algoritmo de Porter comprime el vocabulario del índice por el 43% en el texto inglés.

La lematización parece, en consecuencia, fuertemente dependiente del idioma en que se encuentran documentos y consultas, de manera que resulta difícil aplicar algoritmos diseñados para un idioma a información en otra lengua diferente. No solamente los sufijos y raíces son diferentes, sino que la forma en que aquellos se adhieren a estas es distinta. No obstante, se han propuesto sistemas elementales de lematización que son básicamente independientes del idioma. Éste es el caso de *n-grams* y, en buena medida, de *s-stemmers*, aunque éstos requieren alguna pequeña adaptación.

En cuanto a la lengua persa, fueron construidos algunos lematizadores cuya la mayoría son analizadores lingüísticos utilizando la estructura de palabras y reglas morfológicas de la lengua. La siguiente sección está dedicada a describir todos los trabajos previos en la construcción de algoritmos de lematización.

3.5.1. Lematizador Bon

Bon es uno de los primeros lematizadores desarrollado por el departamento de ingeniería de informática de la universidad tecnología Amir Kabir de Teherán (Tashakori, et al., 2002). Bon, es un lematizador iterativo basado en el sufijo más largo similar al lematizador desarrollado por Lovins (Lovins, 1968). Un lematizador iterativo basado en el sufijo más largo elimina la cadena más larga posible de caracteres de una palabra de acuerdo con un conjunto de reglas. Este proceso se repite hasta que no se pueda quitar más caracteres. Después de que se han eliminados todos los caracteres, puede que la raíz restante no sea correcta. Para controlar este error, Bon utiliza el método de re-codificación. Esta técnica es una transformación de contexto sensible de la forma $AxC \rightarrow AyC$ donde A y C especifican el contexto de la transformación, x es la cadena de entrada e y es la cadena transformada. En el procedimiento de Bon, el afijo

se elimina de acuerdo con un conjunto de las reglas gramaticales de la lengua ya preparado y utilizando tres bases de datos (raíces de la lengua persa, infinitivos y *Mokassar*[14]).

Este lematizador tiene dos procedimientos principales. En cada uno, la palabra se verifica muchas veces buscando en las bases de datos. Para evaluar el lematizador, se considero una pequeña colección de 450 documentos cortos y 32 consultas. Las medidas de precisión y exhaustividad fueron calculadas utilizando un modelo booleano de recuperación. La Tabla 3.12 muestra los valores promedios de precisión y exhaustividad sobre todas las consultas con y sin del uso de lematizador (Tashakori, et al., 2002).

Tabla 3.12: Comparación de la eficacia de recuperación del lematizador Bon

Experimentos	Exhaustividad	Precisión
Sin lematizador	0,3595258	0,8974702
Uso de lematizador	0,5421372	0,8397220

Observaciones

La evaluación fue verificada utilizando una pequeña colección (450 resúmenes de artículos) de documentos que puede no ser muy representativa. Aunque notamos un aumento de la tasa de exhaustividad, hay un ligero descenso de precisión. El lematizador es fuertemente dependiente de base de datos es decir que, desde el principio, para buscar el lema de una palabra hay que mirar en diferentes bases de datos y tal vez al final, el lema no será encontrado. Eso se explica una vez que el lematizador se haya eliminado todos los caracteres, es posible que el lema obtenido no sea siempre correcto porque hay muchas excepciones para hacer reglas de lematización. Así que, el tiempo del proceso de lematización es demasiado largo y sobre todo no se recomienda

[14] Mokassar son aquellos sustantivos en la lengua árabe que tienen forma irregular de plural y algunos de ellos se utilizan en persa.

la aplicación de este algoritmo para los procesos en línea. Además, la implementación del algoritmo depende de la construcción de tres bases de datos, lo que representa un trabajo considerable y los elementos de la base de datos de *Mokassar* no son muy claros. Hay que guardar todas las palabras de *Mokassar* de la lengua árabe o bien las que son más utilizadas en la lengua persa. Un inconveniente a tomar en cuenta es que este lematizador no está disponible al público.

3.5.2. Lematizador de ISRI

El lematizador desarrollado por el instituto de investigaciones en ciencias de información de la universidad de Nevada (Taghva, et al., 2005) es similar al lematizador de Porter (Porter, 1980) ya que ambos están basados en la morfología de la lengua. Ambos lematizadores buscan ciertos sufijos y utilizan las múltiples fases que se conforman con las reglas del apilamiento de sufijo (*suffix stacking*). Además, fijan un límite inferior en el número de caracteres para considerar que se ha conseguido la raíz. Sin embargo, hay diferencias importantes, por ejemplo, el lematizador de Porter identifica patrones de consonantes y de vocales para estimar el contenido de información. En persa, muchas vocales habladas no se escriben, así que el lematizador no puede contarlas. Por lo tanto, el lematizador de ISRI utiliza la longitud mínima para buscar la raíz y el valor de la longitud mínima es igual a tres (Taghva, et al., 2005).

El primer paso del algoritmo es encontrar una sub-cadena terminal de la palabra de entrada que está en la lista de los sufijos. Esta lista de sufijos fue elaborada a mano basada en la gramática de la lengua persa. Si hay varios sufijos en una palabra, el lematizador elige el sufijo más largo que dejaría una raíz con tres o más caracteres. Para determinar qué sufijo termina la palabra de entrada, el lematizador utiliza una máquina de estado finito (*DFA, Deterministic Finite Automata*). Cada estado es un estado de aceptación. La *DFA* se codifica como una matriz bidimensional. Las filas representan los estados y las columnas representan las letras de entrada.

El lematizador fue evaluado utilizando una pequeña colección de documentos (corpus de ISRI citado en el parte de corpus, ver la Sección 3.2.2). Se utilizó modelo

espacio vectorial para observar el efecto del lematizador sobre las valores de precisión y exhaustividad. La evaluación de lematizador fue mediante la comparación entre los valores de precisión y exhaustividad en dos casos. El primero sin aplicar el lematizador y el segundo con la aplicación del lematizador y también con la eliminación de las palabras vacías. Recordamos que una lista de 155 palabras vacías fue identificada por el *ISRI* (ver la Sección 3.4.1). Para ver mejor el rendimiento del lematizador ponemos en la Figura 3.5 la curva de precisión-exhaustividad según los resultados obtenidos (Taghva, et al., 2005).

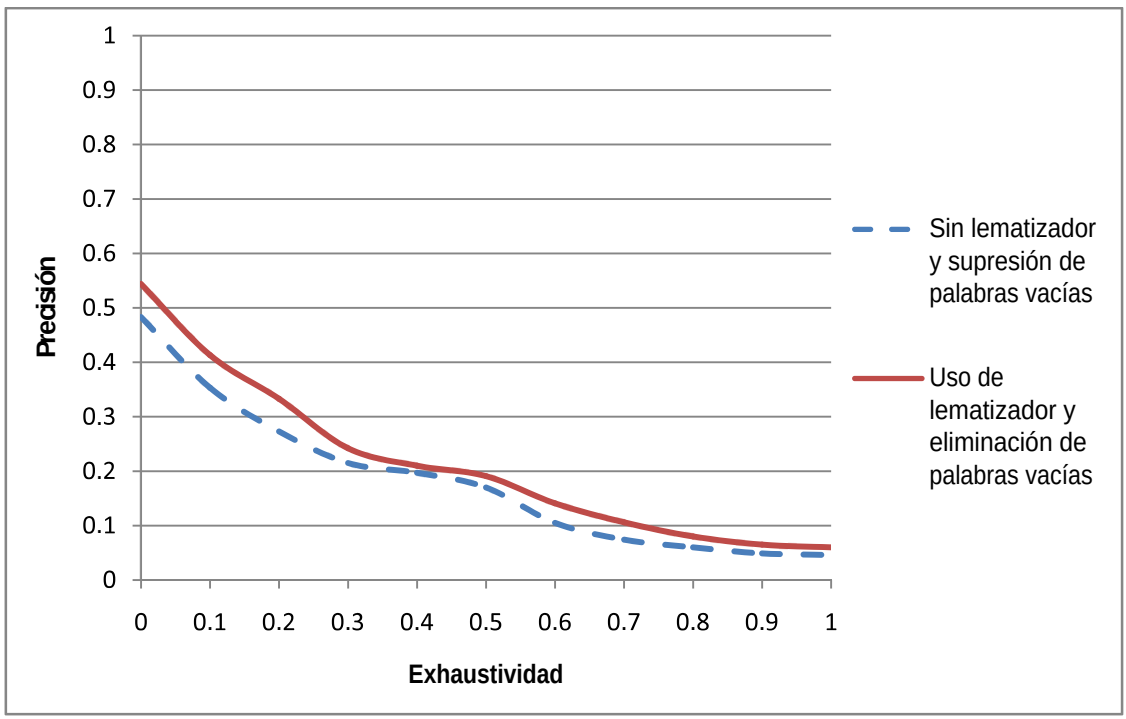

Figura 3.5: Curva de precisión-exhaustividad del lematizador de ISRI

Observaciones

Los resultados obtenidos revelan que el uso de lematizador mejora la recuperación y la media de precisión aumenta alrededor de 18%. Este algoritmo no considera los prefijos ni todos los signos plurales de la lengua persa. Los plurales irregulares (*Mokassar*) tampoco están considerados en la implementación del algoritmo. La longitud mínima de la raíz es igual a 3, pero algunas palabras en persa pueden tener sólo dos letras. Por

ejemplo, la palabra حد tiene dos caracteres y significa “limite” en español. Parece que la lista de los sufijos no está completa y hay que elaborarla considerando todos los sufijos posibles. De nuevo, este lematizador tampoco está disponible al público.

3.5.3. Perstem

Perstem es un lematizador y analizador morfológico ligero desarrollado por Jon Dehdari del departamento de lingüística de la universidad del estado de Ohio (Jadidinejad, et al., 2009). Está escrito en Perl y utiliza una serie de sustituciones de expresiones regulares para separar morfemas flexivos de la raíz. La palabra entrada se asigna isomórficamente a un formato interno de romanización para el rendimiento y la consistencia interna. En primero, la palabra se busca en una tabla de dispersión (*hash table*) y si la palabra corresponde a una clave entonces la salida es su valor asociado y no se producen sustituciones de expresiones regulares. Este paso preliminar de la utilización de una tabla *hash* sirve para múltiples propósitos. El primer propósito es acelerar el tratamiento de las palabras muy comunes en el texto. El valor asociado a estas palabras es una cadena vacía. Otro objetivo de la tabla *hash* es ayudar a lematizar las palabras plurales irregulares. El tercer objetivo es eliminar las palabras vacías que tienen una cadena vacía como el valor asociado. El propósito final de la tabla hash es corregir algunas palabras que han conseguido una incorrecta raíz por las expresiones regulares. El lematizador Perstem tiene actualmente cerca de 50 reglas de sustituciones de expresiones regulares.

Con el fin de investigar la efectividad de Perstem, se llevó a cabo dos experimentos diferentes utilizando el mismo modelo de recuperación. En primer lugar, fue indexada y recuperada una parte del corpus Hamshahri, utilizando el buscador Indri (Metzler, et al., 2004) sin aplicar Perstem. En segundo lugar, se aplicó Perstem a todos los documentos y consultas y se creó unos nuevos corpus y consultas lematizados. Después, El mismo método de indexación y de recuperación utilizando Indri fue aplicado sobre el corpus lematizado. La Tabla 3.13 compara la eficacia de estos experimentos en CLEF2009 (Jadidinejad, et al., 2009) (Ferro, et al., 2009).

Tabla 3.13: Comparación de rendimiento de Perstem

Experimentos	Relevantes - Recuperados	Exhaustividad	*MAP*
Sin Perstem	2.670 / 4.464	0,5981	0,1964
Uso de Perstem	3.820 / 4.464	0,8557	0,3762

Observaciones

Al contrario de lo que sucede con los anteriores, este lematizador está disponible al público y se puede descargar desde la página Web del autor[15]. La evaluación de Perstem fue realizada sobre 23.536 documentos agrupados donde hay 19.072 documentos no relevantes y 4.464 documentos relevantes utilizando 50 consultas. Las comparaciones entre los dos experimentos muestran que Perstem mejora tanto la precisión como la exhaustividad. Hay un aumento del 43% en la exhaustividad y 91% en la medida promedio de precisión (*MAP*). Debido a las muchas excepciones en la gramática persa, parece que las 50 reglas de sustituciones de expresiones regulares no son suficientes y se tendrán que añadir más reglas gramaticales a los algoritmos y también considerar las palabras ambiguas en la construcción de tablas *hash*. Hemos verificado el código fuente del lematizador y constatamos que podemos mejorar la efectividad de Perstem efectuando unas modificaciones. Por ejemplo, considerando algunos signos de sufijos plurales árabes que se utilizan por lo general en la lengua persa.

3.5.4. Lematizador de Estahbanati

Otro algoritmo de lematización fue propuesto por Estahbanati (Estahbanati, et al., 2011). Este lematizador es el mismo que el de ISRI (ver la Sección 3.5.2) con la diferencia de que se utiliza una base de datos para guardar las excepciones con el fin de

[15] http://www.ling.ohio-state.edu/~jonsafari/

disminuir la tasa de error. La función de la base de datos es mejorar el lematizador cuando se producen las siguientes excepciones. En primer lugar, hay algunas palabras que son estructuralmente similares a otras y que no deben ser utilizadas como prefijos o sufijos de lematizador. En segundo lugar, hay algunas palabras en plural llamado *Mokassar*, o plurales irregulares, para las cuales no existen normas para determinar sus raíces. En tercer lugar, la base de datos contiene las palabras que tiene menos de tres letras. La implementación del algoritmo es también la misma que ISRI y se utiliza una máquina de estados finitos donde la aplicación incluye un lematizador de sufijos y otro de prefijos. El lematizador de sufijos tiene 15 estados y mientras que el lematizador de prefijos tiene sólo 2.

El método de recuento de errores se utiliza para evaluar el rendimiento del algoritmo. La colección de prueba contenía sólo 5 documentos de internet con 927 palabras. En la Tabla 3.14 podemos ver los resultados de esta primera fase de experimentos que no se utiliza la base de datos (Estahbanati, et al., 2011).

Tabla 3.14: Resultados del primer algoritmo de Estahbanati

N° de prueba	Total palabras	Resultados correctos	Resultados incorrectos	Porcentaje de resultados correctos
1	41	33	8	80,50
2	89	76	13	85,40
3	120	97	23	80,87
4	130	108	22	83,07
5	547	492	55	89,94

En la segunda fase de experimentos se aplica también la base de datos en el funcionamiento del algoritmo y utilizando el método de recuento de errores sobre 4 documentos. En la Figura 3.6 tenemos la comparación entre los dos algoritmos (Estahbanati, et al., 2011).

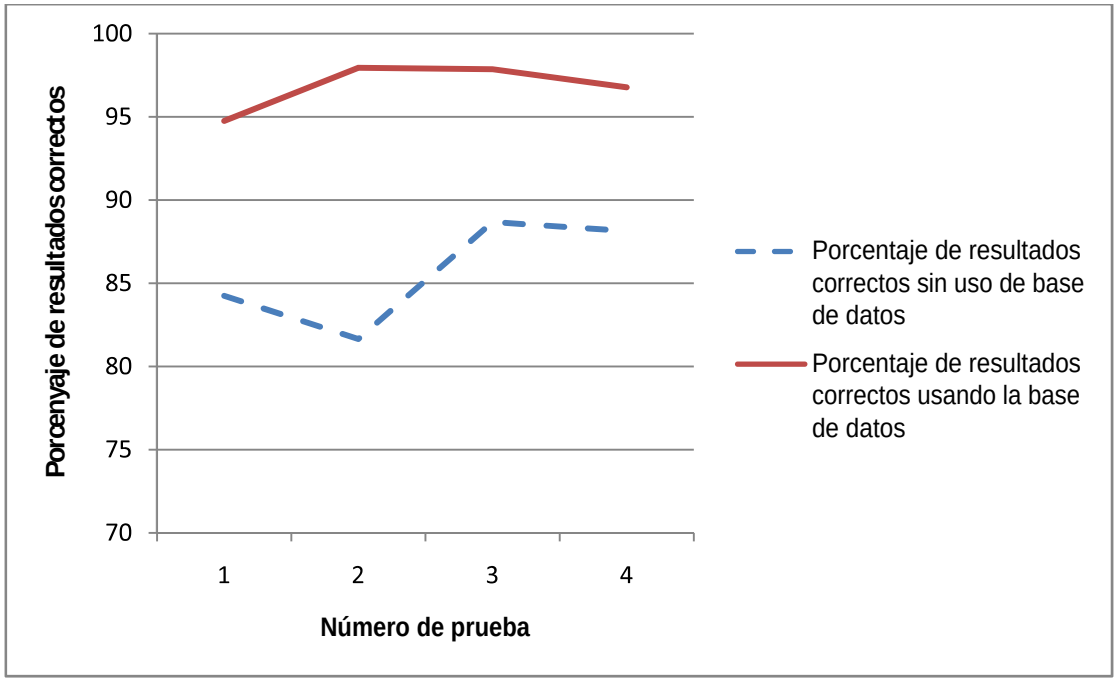

Figura 3.6: Comparación de dos algoritmos de Estahbanati

Observaciones

Aunque es posible evaluar el lematizador contando los números errores que se producen durante el proceso de lematización, en el ámbito de recuperación de información es más significativo de proceder a las medidas de precisión y exhaustividad de las consultas para ver el comportamiento del lematizador con los documentos. Aún suponiendo el método de evaluación de recuento de errores, hay que probar con mucho más documentos para poder llegar a una conclusión. Admitiendo los resultados obtenidos, el promedio de los resultados correctos sin utilizar la base de datos es 85,68% y el uso de la base de datos es 96,83%. Podría ser que los errores, en el primer algoritmo, se deban a las palabras no verbales que son estructuralmente similares a los verbos o debido a las palabras cuyos lemas son de menos de tres letras. El hecho de utilizar la base de datos se nota ya que la tasa de errores disminuye, pero de este modo, el algoritmo depende de la base de datos. Los errores, en este caso, pueden ser debidos a que el analizador lingüístico no puede detectar el tipo correcto de las palabras.

3.5.5. Lematizador de SECE

Otro algoritmo de eliminación de afijos fue propuesto por SECE (*School of Electrical and Computer Engineering*) de la universidad de Teherán (Rahimtoroghi, et al., 2010). Este lematizador utiliza la estructura de palabras y reglas morfológicas del idioma para reconocer el lema de una palabra. Se compone de 33 reglas para la descripción del algoritmo de lematizador donde las reglas están escritas sobre la base de la morfología de la lengua y la estructura de la derivación de palabras. Este algoritmo sólo se ocupa de eliminar los sufijos y trata los sustantivos, adjetivos y adverbios. Los autores piensan que los sufijos son más predominantes que otras formas en la morfología de la lengua persa y que la mayoría de las consultas en un sistema de RI sólo contienen nombres, adjetivos y adverbios y no verbos.

La evaluación se hizo mediante el corpus Hamshahri utilizando 50 consultas y el modelo probabilístico de recuperación de información Okapi BM25 (Robertson, et al., 2000). Los experimentos muestran una reducción del tamaño de índice por el factor de 6% y las medidas de precisión y exhaustividad están en la siguiente Tabla 3.15 (Rahimtoroghi, et al., 2010).

Tabla 3.15: Resultados obtenidos por el lematizador de SECE

Medidas	Sistema de RI sin lematizador	Sistema de RI con lematizador
MAP	0,4031	0,4224
Exhaustividad	0,8592	0,8656

Observaciones

Los resultados demuestran que la aplicación del lematizador aumenta el promedio medio de precisión en un 4,78% y la exhaustividad el 0,74%. Estos resultados parecen ser bajos porque el algoritmo de lematizador solo tiene 33 reglas de lematización y no trata los verbos. Hay que considerar mucho más reglas gramaticales y también las excepciones para la construcción de lematizador en persa. Otro punto débil del lematizador es que no considera los prefijos.

3.5.6. Resumen

La mayor parte de los lematizadores que hemos analizados están basados en reglas similares al algoritmo de Porter. Este enfoque de algoritmo elimina el prefijo o sufijo de acuerdo con un conjunto de reglas gramaticales. Hay algunos algoritmos que son lematizadores ligeros que tratan sólo los sufijos y los signos plurales. Hay otros que son más completos que tratan también los verbos y utilizan una base de datos para las excepciones. La mayoría de estos lematizadores fueron evaluados usando un subconjunto limitado de palabras o un corpus limitado y los métodos de evaluación eran diferentes. Desgraciadamente, no tenemos acceso a las fuentes de todos estos algoritmos para poder comprobarlos sobre la misma colección de documentos.

La eficacia de un algoritmo de lematización se debe evaluar en un sistema de IR utilizando las medidas estándares (como la precisión y la exhaustividad). A continuación, detallamos en la Tabla 3.16 las ventajas y las limitaciones de cada algoritmo. Los resultados que tenemos no pueden ser considerados en una comparación absoluta, sino más bien como una comparación relativa.

Tabla 3.16: Comparación entre diferentes lematizadores persas

Algoritmos de lematización	Ventajas	Limitaciones
Bon	sufijos y prefijos plurales irregulares verbos	alto tiempo de proceso construcción de 3 bases de datos
ISRI	sufijos	no manejar plurales irregulares no manejar todos signos de plural no quita el prefijo no detectar el lema con 2 letras
Perstem	sufijos y prefijos	no manejar todos signos de plural construcción de tablas internas
Estahbanati	sufijos y prefijos	no todos prefijos alto tiempo de proceso no detectar todo tipo de palabras construcción de bases de datos de las excepciones
SECE	sufijos	no trata los prefijos no trata los verbos

3.6. Los Modelos de Recuperación de Información

El objetivo de cualquier sistema de RI es la necesidad de determinar qué documentos son relevantes y cuáles no, siendo necesario un algoritmo que realice la clasificación de los documentos. El algoritmo actuará de acuerdo a una serie de premisas para evaluar qué es relevante y qué no lo es, de manera que diferentes conjuntos de premisas determinan diferentes modelos de recuperación de información. La diferencia fundamental entre los distintos modelos de recuperación de información existentes se basa en la forma en que se define e implementa en ellos el concepto de relevancia. Es necesario establecer de antemano las premisas que debe cumplir un documento para ser relevante para una consulta dada. Distintos conjuntos de premisas nos proporcionan distintos modelos de recuperación de información.

Los tres modelos clásicos de IR son el modelo booleano, vectorial y probabilístico (Baeza-Yates, et al., 1999). En el modelo booleano los documentos y las consultas están representadas por conjuntos de términos índice (Van-Rijsbergen, 1979). En el modelo vectorial, los documentos y las consultas están representados como vectores en un espacio t-dimensional (Salton, et al., 1988). Por ello, el modelo se denomina algebraico. En el modelo probabilístico, la herramienta para la modelización de los documentos y la consulta se basan en la teoría de la probabilidad (Robertson, et al., 1976). Los modelos de recuperación de información han ido evolucionado partiendo de teorías iníciales ya han sido extendido para lograr una mejor eficiencia y precisión. Un modelo de RI es evaluado de acuerdo a su capacidad para identificar los documentos relevantes y eliminar de la respuesta aquellos que no lo son. La proporción de documentos que son recuperados frente al número de documentos que se cree que son relevantes se llama eficiencia y, mientras que el factor de pertinencia o precisión se basa en la porción de documentos recuperados que son realmente relevantes (Swanson, 1988).

Baeza-Yates (Baeza-Yates, et al., 1999) distingue principalmente dos divisiones importantes dentro de dichos modelos para realizar la tarea de recuperación *Adhoc* y filtrado de documentos, que son: los modelos clásicos y los modelos estructurados. A los modelos clásicos representan los tres modelos booleano,

vectorial y probabilístico que son la base muchos otros. Entre los modelos estructurados encontramos las listas no coincidentes y el modelo de nodos próximos. La Figura 3.7 muestra gráficamente una taxonomía completa de los modelos como de sus extensiones.

En esta última parte del capítulo se describe los trabajos previos sobre la aplicación de los modelos de RI en documentos persas. De hecho, todos estos modelos son aquellos que se han creados a lo largo del tiempo para otros idiomas como el inglés.

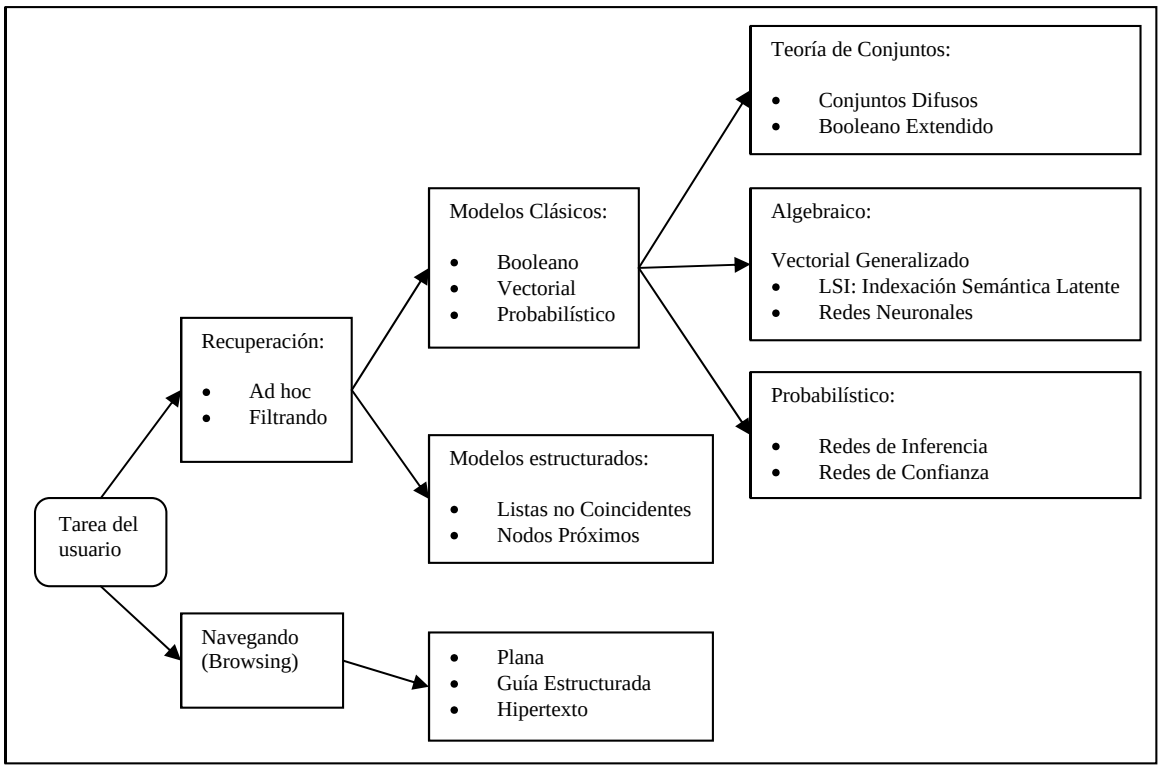

Figura 3.7: Taxonomía de los modelos de RI (adaptada de Baeza-Yates y Ribeiro-Neto, 1999)

3.6.1. Modelo Difuso

El sistema de RI en persa basando en la lógica difusa (*Fuzzy logic*) llamando "FuFaIR" fue realizado por Nayyeri (Nayyeri, et al., 2006) del departamento de eléctrica y informática de la universidad de Teherán. En 1965, Lotfi A. Zadeh (Zadeh, 1965), un matemático estadounidense de origen iraní, publicó un documento en donde por primera vez se menciona el término "*Fuzzy Logic*", conocida en la lengua español

como "lógica difusa". En recuperación de información se ha propuesto el método de RI difusa basado en la teoría de conjuntos difusos para mejorar la desventaja del modelo lógico booleano que no puede manejar información ambigua e imprecisa. Un conjunto difuso, es un conjunto que puede contener elementos de forma parcial. Es decir que la propiedad $x \in A$ puede ser cierta con un grado de verdad. Se mide esta posibilidad de pertenecer (o pertenencia) con un número $\mu_A(x)$ entre 0 y 1, llamado grado de pertenencia de *x* a *A*. Si es 0, *x* no pertenece a *A*, si es 1 entonces $x \in A$ completamente y si $0 < \mu_A(x) < 1$, *x* pertenece a *A* de una manera parcial. Un subconjunto *A* de *B* se caracteriza, por tanto, por esta función de pertenencia μ_A, de *B* en [0,1]. Es preciso fijar el conjunto *B* para definir la función μ_A que a su vez define *A*. Por eso se habla de subconjunto difuso y no de conjunto difuso. Nótese que μ_A es una proposición en el contexto de la lógica difusa, y no de la lógica usual binaria, que sólo admite dos valores: cierto o falso. La recuperación de información implica considerar dos conjuntos finitos, un conjunto de los términos de índice reconocidos $X = \{x_1, x_2, \dots, x_n\}$ y un conjunto de documentos relevantes $Y = \{y_1, y_2, \dots, y_m\}$.

En la recuperación de información difusa, la relevancia de los términos de índice a los documentos individuales se expresa mediante una relación difusa $R = X \times Y \rightarrow \{0, 1\}$ de tal manera que el valor de pertenencia (x_i, y_j) específica para cada $x_i \in X$ y $y_j \in Y$ el grado de relevancia del término de índice x_i al documento y_j. El valor de pertenencia aún se puede calcular fácilmente con parámetros clásicos de RI como *tf /idf* (frecuencia del término/ frecuencia inversa del documento). En este trabajo, se ha utilizado la siguiente fórmula para el cálculo el valor de pertenencia:

$$\mu_t(d) = \frac{f_{t,d}}{max_{t_k}(f_{t_k,d})} \times \frac{idf(t)}{max_{t_k}(idf(t))} \qquad (3.1)$$

donde $\mu_t(d)$ es el valor de pertenencia de documento d en el conjunto difuso del término t. $f_{t,d}$ representa la frecuencia del término t en el documento t. $max_{t_k}(f_{t_k,d})$ es la frecuencia máxima para cualquier término en el documento d. *idf* (*t*) o la frecuencia inversa del documento representa la porción de la colección que contiene el término t.

Eso se calcula como $\log(\frac{N}{n}) + 1$, donde N es el número de documentos en la colección y n es el número de documentos con el término t.

Para evaluar el rendimiento del sistema, una comparación fue establecida entre este modelo y el modelo clásico de espacio vectorial con esquema de ponderación *Lnu.ltu* (Salton, et al., 1986). El sistema de ponderación *Lnu.ltu* es uno de los más eficaces sistemas de ponderación para el modelo de espacio vectorial. En el esquema de ponderación *Lnu.ltu*, los documentos se ponderan con *Lnu* y la consulta del usuario se pondera con *ltu* que se calcula como sigue:

$$Lnu = \frac{\frac{1 + \log(tf)}{1 + \log\big(average(tf)\big)}}{(slope \times N.U.T) + (1 - slope) \times pivot} \qquad (3.2)$$

$$ltu = \frac{\ln(tf) + 1.0 \times \ln\frac{N}{n}}{(slope \times N.U.T) + (1 - slope) \times pivot} \qquad (3.3)$$

tf es la frecuencia del términos, *N* es el número de los documentos en toda la colección, *n* es el número de documentos en los que ocurre el término y $N.U.T$ se el número de términos únicos dentro del documento especificado. *Slope* y *pivot* son variables constantes que se utilizan en este método y se llaman normalización pivotada. El pivote constante es el número promedio de términos singulares en el conjunto de colección.

La evaluación de este modelo se hizo utilizando los documentos de la colección de Hamshahri. En la Figura 3.8 se muestra los resultados obtenidos aplicando el modelo difuso y espacio vectorial (Nayyeri, et al., 2006).

Observaciones

Para la evaluación del modelo, se han considerado sólo 30 consultas y la medida de la precisión sobre los 20 primeros documentos recuperados sin tener en cuenta del valor de exhaustividad. Es preferible considerar más consultas y medir los valores de precisión y exhaustividad sobre todo el conjunto de documentos recuperados para realizar un juicio sobre el rendimiento del modelo. Pero basándose sobre la

comparación de estos resultados obtenidos, podemos decir que el rendimiento del modelo difuso aplicado a los documentos persas es mejor que el modelo de espacio vectorial.

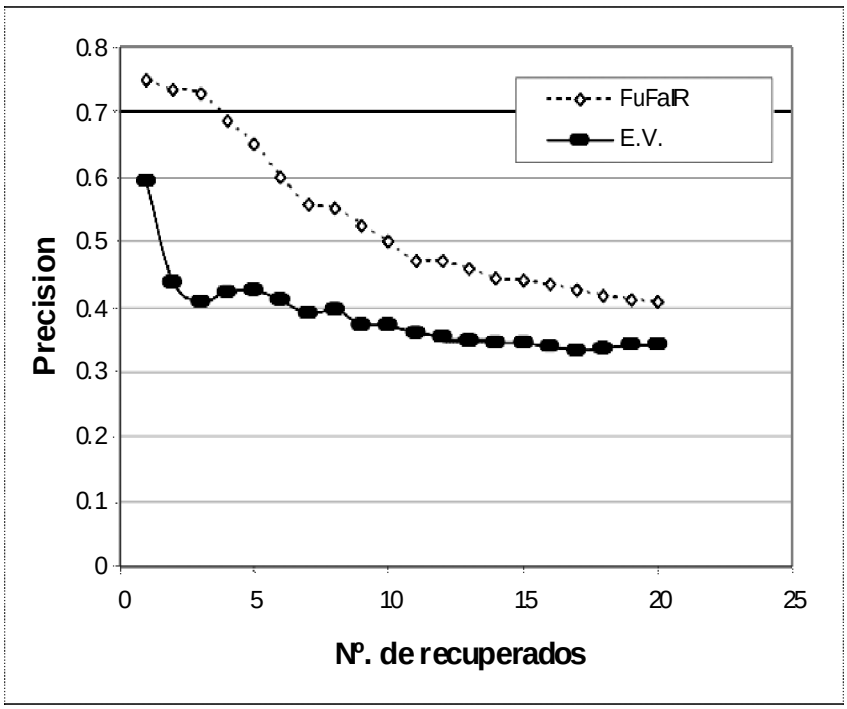

Figura 3.8: Precisión del modelo FuFaIR en contra de espacio vectorial, tomado de (Nayyeri, et al., 2006)

3.6.2. Modelo del Lenguaje

Otro modelo aplicado a la recuperación de información en persa se basa en la aplicación de técnicas de los modelos de lenguaje propuesto por *ISRI* de la universidad de Nevada (Taghva, et al., 2004). El modelo de lenguaje estadístico se utilizó por primera vez por Andrei Markov[16] para modelar secuencias de letras en la lengua rusa. Los modelos de lenguaje ya se han aplicado con éxito para el procesamiento del lenguaje natural, reconocimiento automático de habla y RI, así como a muchos otros ámbitos (Hiemstra, 2000).

[16] Andréi Andréyevich Márkov (14 de junio de 1856 - 20 de julio de 1922) fue un matemático ruso conocido por sus trabajos en la teoría de los números y la teoría de probabilidades.

Una manera de entender el modelado del lenguaje estadístico es concebir el problema de recuperación de documentos en términos de la teoría de información (Zhai, et al., 2001) (Chen, 1996). De acuerdo con este punto de vista, el proceso de recuperación equivale a la decodificación de un documento transmitido a través de un flujo de información como se ilustra en la siguiente Figura 3.9.

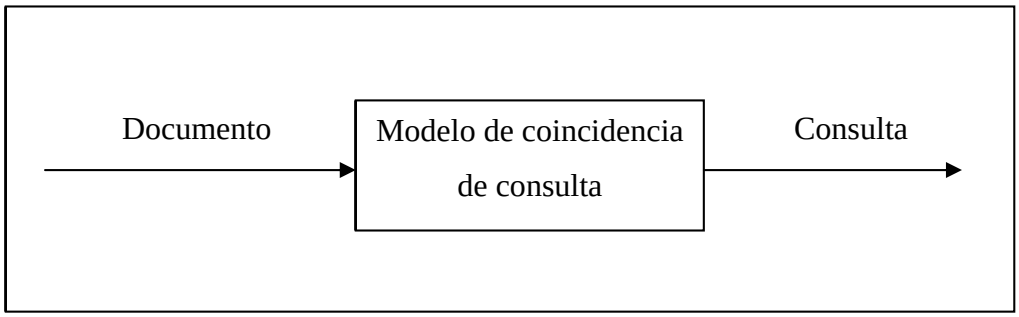

Figura 3.9: Recuperación según la teoría de información

En el caso de la recuperación de información, el canal de información ruidoso es la intención del usuario que trata de imaginar qué documento desea el usuario (Miller, et al., 1999). La consulta que el usuario formula se convierte en la única evidencia de lo que es el mensaje original (documento). En este enfoque, el sistema de recuperación es una función de asignación de los términos de la consulta al documento de forma que esta función intenta decodificar los términos de la consulta en el documento original. La función óptima recuperará el documento con la mayor probabilidad tiendo en cuenta los términos de la consulta. La función óptima utilizada en esta evaluación es la que propuso Djored Hiemstra (Hiemstra, 2000) en su trabajo de tesis titulado "*Using Language Models for Information Retrieval*" como:

$$HLM4(d) = \log(\sum \mathrm{tf(t,d)}) + \sum_{i=1}^{n} \log\left(1 + \frac{\lambda_i tf(t_i,d)(\sum_t df(t))}{(1-\lambda_i)df(t_i)(\sum_t tf(t,d))}\right) \quad (3.4)$$

donde *tf* (t_i,*d*) es el numero de ocurrencia del término t_i en el documento *d* que es la frecuencia del término t_i y *df* (t_i) es el número de documentos en la colección dentro de los cuales aparecen el termino t_i. Este valor se llamaría generalmente la frecuencia de documento de término t_i. λ_i es el peso dado a los términos importantes y 1- λ_i es el peso

de términos sin importancia relativa al documento D_i. *HLM4* es la cuarta versión de la ecuación del modelo de lenguaje por Hiemstra (Hiemstra, 2000).

La evaluación se hizo por Taghva et al. (Taghva, et al., 2004) comparando este modelo de lenguaje con una implementación típica del modelo de espacio vectorial usando la medida de similitud del coseno (Witten, et al., 1999). La colección de prueba fue creada por ISIR con 1.647 documentos y 60 consultas (citado en el parte de corpus, ver la Sección 3.4.2). La lista de palabras vacías y el lematizador construidos por ISIR fueron también considerados para completar la evaluación. Con unos experimentos en la colección, el valor óptimo del parámetro λ fue fijado como λ=0,035 en el caso de eliminar las palabras vacías y aplicar el lematizador en los documentos de la colección. El valor λ= 0,0485 fue considerado sin usar el lematizador ni quitar las palabras vacías. En la Tabla 3.17 se resume los resultados obtenidos de cuatro pruebas indicando el valor de la precisión media (PVL=eliminación de palabras vacías y aplicación del lematizador; NPVL= no eliminación de palabras vacías y no lematización).

Tabla 3.17: Comparación de precisión media en los casos de coseno y HLM4

Modelo Coseno - NPVL	Coseno - PVL	Modelo HLM4 - NPVL	HLM4 - PVL
0 ,180	0,211	0,220	0,234

Observaciones

Para comprender mejor la comparación entre los diferentes modelos, hemos tomados los resultados citado en la Sección 3.5.2 (el caso del modelo coseno sin lematizador y con palabras vacías) y hemos dibujado las curvas precisión-exhaustividad en la Figura 3.10.

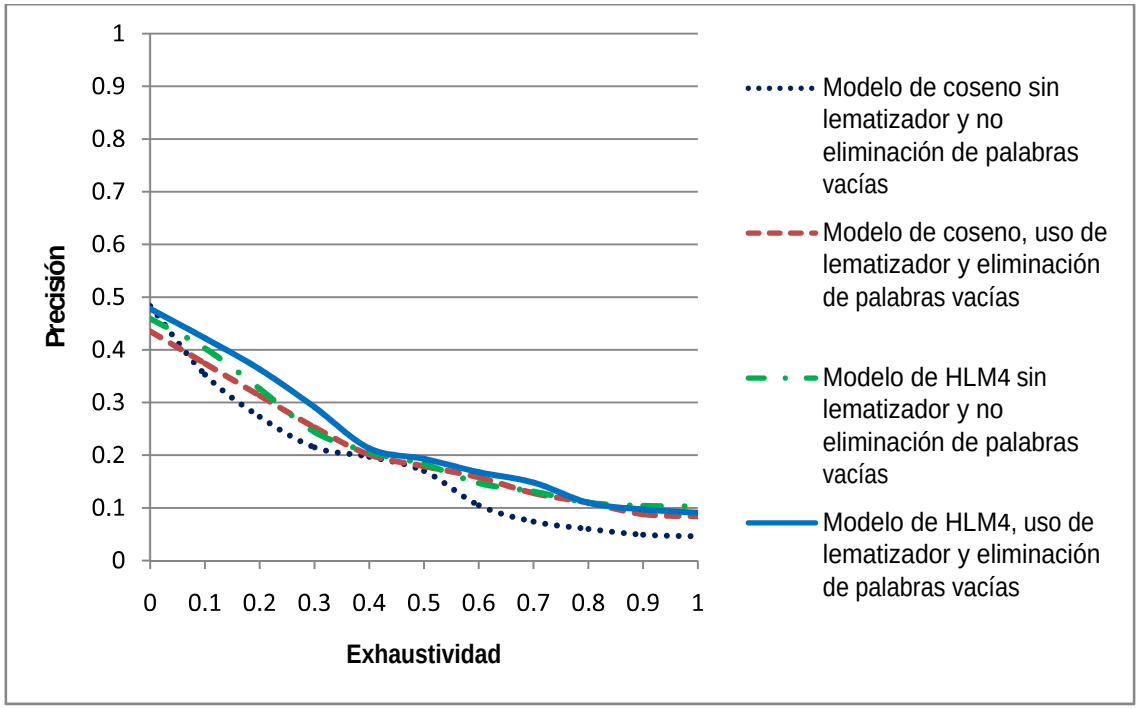

Figura 3.10: Comparación entre modelos de vector coseno y lenguaje *HLM4*

La comparación entre estos cuatros modelos revela que el peor enfoque para recuperar documentos persas es el modelo de espacio vectorial sin aplicar la lematización ni eliminar las palabras vacías lo que parecería obvio. Los mejores resultados se obtuvieron mediante el uso de lematizador y la eliminación de palabras vacías aplicando el modelo del lenguaje. En general, el enfoque de modelo de lenguaje mejora la precisión en un promedio de casi 10%. El método *HLM4* con el uso de lematizador y eliminación de palabras vacías mejora la recuperación en un promedio de alrededor del 3% en comparación con el modelo *HLM4* sin lematizador y no eliminación de palabras vacías. Estos experimentos se realizaron con una pequeña colección de documentos no estándar. Sería interesante comprobar estos modelos con un corpus más grande y estándar para llegar a conclusiones significativas.

3.6.3. Modelos N-gramas y Análisis del Contexto Local

Otros experimentos fueron realizados por Oroumchian et al. (Oroumchian, et al., 2007) de la universidad de Teherán, aplicando los métodos N-gramas y análisis del contexto local. El análisis del contexto local (*LCA, Local Context Analysis*) es un método de expansión automática de consulta que combina el análisis global y la realimentación local y fue presentado por Xu (Xu, et al., 1996). La expansión de consulta es un proceso de reformulación automática del sistema que permite añadir nuevos términos a la consulta para mejorar el contexto de la consulta original del usuario. Esto se consigue mediante procesos de agrupación (*clustering*), que determinan la frecuencia de aparición de un grupo de términos contiguos, relacionados con la consulta del usuario, presentes en documentos clasificados dentro de un mismo ámbito temático (en el caso de análisis del contexto local) y en torno a toda la colección (en el caso de análisis del contexto global). Análisis del contexto local se lleva a cabo en tres pasos: En primer lugar, se ejecuta la consulta original y se recupera los *n* mejores documentos de respuesta a una consulta. Estos documentos son divididos en pasajes fijos o ventanas de texto y se clasifican estos pasajes como si fueran documentos. En segundo lugar, para cada concepto *c* dentro de los pasajes mejor evaluados se calcula la similitud *sim*(*q*,*c*) entre toda la consulta *q* y el concepto *c* usando una variación del *ranking tf-idf*. La función de similitud se calcula como:

$$sim(q,c) = \prod_{k_i \in q} (\delta + \frac{\log(f(c,k_i) \times idf_c)}{\log n})^{idf_i} \qquad (3.5)$$

δ es un valor constante pequeño (cerca de 0,1) y la función f (c,k_i) se obtiene como:

$$f(c,k_i) = \sum_{j=1}^{n} Pf_{i,j} \times Pf_{c,j} \qquad (3.6)$$

$Pf_{i,j}$ es la frecuencia del término *i* en el *j-ésimo* pasaje; k_i en el *j-ésimo*pasaje y $Pf_{c,j}$ es la frecuencia del concepto *c* en el *j-ésimo* pasaje.

En la ecuación 3.5, se pueden calcular los factores de la frecuencia inversa del documento como:

$$idf_i = \max(1, \frac{\log_{10}(N/np_i)}{5}) \, , idf_c = \max(1, \frac{\log_{10}(N/np_c}{5}) \qquad (3.7)$$

donde N es el numero de pasajes, np_i es el número de pasajes que contienen el término k_i y np_c es el número de pasajes que contienen el concepto *c*. En tercer lugar, después de estos cálculos, la *m* parte superior de los conceptos clasificados se añaden a la consulta original y el método de recuperación inicial se realiza con la consulta expandida. Con excepción de que ahora la consulta expandida es ponderada. Los términos de la consulta original tienen un peso de 2 y los conceptos añadidos se clasifican como 1-(0,9× i)/m, en el que *i* es el rango del concepto en los conceptos clasificados.

N-gramas son cadenas de longitud N generadas a partir de palabras del texto. En los enfoques tradicionales del modelo espacio vectorial, las dimensiones del espacio de documento para una colección dada de documentos son palabras o a veces frases que ocurren en la colección. Por el contrario, en el enfoque de N-gramas las dimensiones del espacio de documento son N-gramas, es decir, cadenas de n caracteres consecutivos extraídos de palabras. Dado que el número de posibles cadenas de longitud N (para gran valor de N) es mucho menor que el número de posibles palabras individuales en un idioma, por lo tanto, los enfoques N-gramas tienen menor dimensionalidad (Cavnar, 1995). Así, el método de N-gramas es un enfoque puramente estadístico que mide propiedades estadísticas de las cadenas del texto en una colección sin tener en cuenta el vocabulario y las propiedades léxicas o semánticas de lenguaje natural en que se escriben los documentos.

En el caso de estudio de las N-garmas aplicadas al persa, la evaluación fue realizada por Oroumchian et al. (Oroumchian, et al., 2007) con documentos de la colección Hamshahri comparando el modelo espacio vectorial con el modelo de análisis del contexto local. Al principio, se aplica el modelo espacio vectorial eligiendo la ponderación *Lnu.ltu* que hemos descrita en las ecuaciones 3.2 y 3.3. Los resultados muestran que, tomando el valor constante de *Slope*=0,25 y utilizando *P.U.N* (*Pivot Unique Normalization*) consiguen mejores resultados que lo mismo modelo con el valor constante de *Slope* = 0,75 y utilizando *P.C.N* (*Pivot Cosine Normalization*). Esta constatación es consistente con lo que se reporta para el texto inglés (Singhal, et al.,

1996). Con el fin de mejorar los resultados, se aplica el modelo de análisis del contexto local con el mejor sistema (es decir, tomando Lnu.ltu con *Slope* = 0,25 y utilizando P.N.U). LCA se realiza con la expansión de la consulta original del usuario y la ponderación de los términos utilizando la ecuación 3.5. El número de pasajes recuperados fue elegido como 20 y el tamaño de cada pasaje era 300 palabras. El número de términos de consulta expandida fue de 10 lo que significa que se elige 10 principales conceptos clasificados y se añade a la consulta. La Figura 3.11 muestra los valores de precisión y exhaustividad para los tres métodos de recuperación que acabamos de describir.

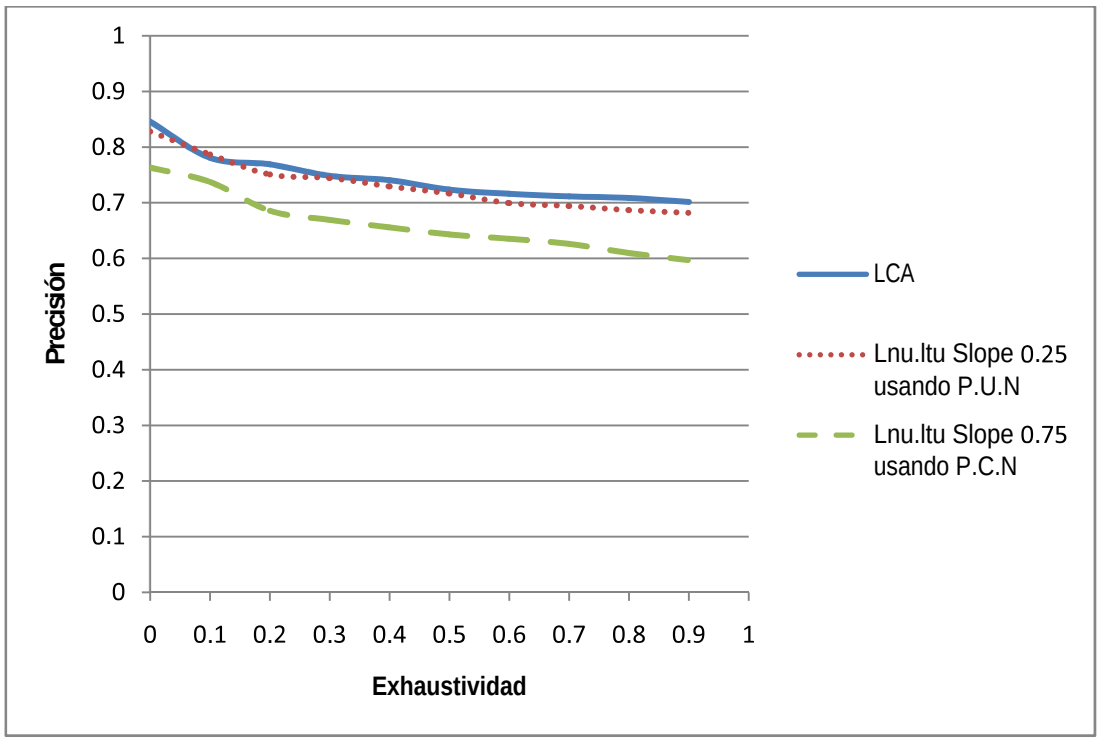

Figura 3.11: Curva precisión-exhaustividad comparando modelos *LCA* y *Lnu.ltu*

La segunda evaluación fue la comprobación de N-grama basado en el modelo espacio vectorial por N=3 y N=4 usando dos sistemas de ponderación uno *Lnu.ltu* y otro *atc.atc* (Salton, et al., 1986). En el caso de sistema de ponderación atc.atc, tanto la consulta del usuario como los documentos son ponderados con *atc* y se calcula de la siguiente manera:

$$atc = 0.5 + 0.5 \times \frac{tf}{\max tf} \times \ln\frac{N}{n} \times \frac{1}{\sqrt{\sum_i w_i^2}} \qquad (3.8)$$

donde N es el número de documentos en toda la colección, n es el número de documentos que contienen i-ésimo término y w_i es $tf \times idf$ para i-ésimo término en cada documento. La Figura 3.12 representa las curvas precisión-exhaustividad deducidas de cada caso descrito (Oroumchian, et al., 2007).

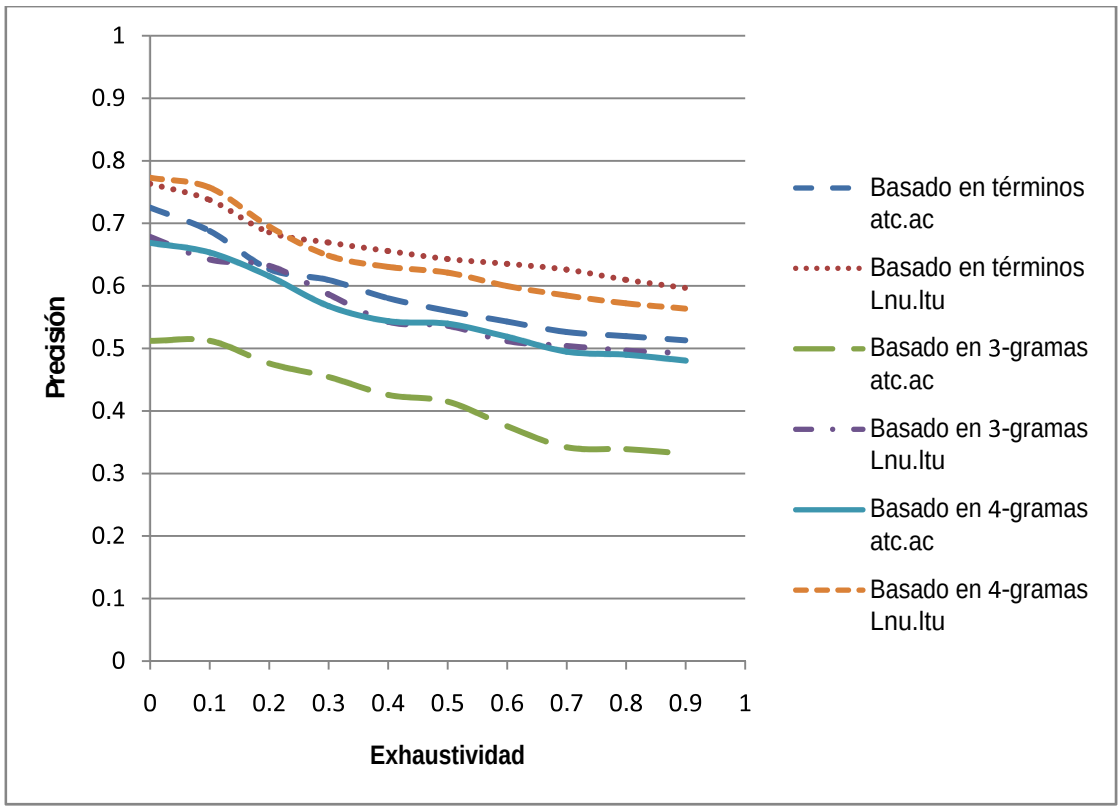

Figura 3.12: Curva precisión-exhaustividad comparando modelos N-gramas

Observaciones

Aunque el corpus Hamshahri fue utilizado para evaluar todos estos modelos sólo se consideró 60 consultas y se evaluaron los 20 primeros documentos recuperados por cada consulta. Analizando los resultados de la primera evaluación, podemos concluir que el modelo de análisis del contexto local mejora marginalmente la precisión (2-3%)

en comparación con el modelo de espacio vectorial. Esto podría ser debido al hecho de que el método de ponderación *Lnu.ltu* está funcionando bien en el idioma persa. Otra explicación también podría ser que los parámetros utilizando por el método de *LCA* son los mismos que se utilizan en *TREC* y puede ser que necesite una modificación para la colección de Hamshahri. En la segunda evaluación, desde los resultados de la experiencia que se muestran por curvas de la Figura 3.12, podemos decir que 4-gramas basados en el espacio vectorial con esquema de ponderación *Lnu.ltu* supera a 3-gramas incluso la configuración basada en los términos y que esquema de ponderación *Lnu.ltu* tienen considerablemente mejor rendimiento que *atc.atc*.

3.6.4. Efectividad de la Recuperación con la Lengua Persa

Como parte de la investigación sobre la lengua persa por el instituto de informática de la universidad de *Neuchâtel*[17], hay un trabajo realizado para estudiar la efectividad de recuperación con el idioma persa (Akasereh, et al., 2012). En realidad, este trabajo se basa en la aplicación de los modelos de RI más comunes sobre documentos persas. La evaluación se hizo utilizando el corpus Hamshahri con 100 consultas (desde *topics* # 551 a # 650 en la colección, donde se toman y evalúan los primeros 50 temas de *CLEF2008* y los últimos 50 de *CLEF2009*). La lista de palabras vacías para estas experiencias es la lista de palabras más frecuentes en la colección con 881 términos. Para la lematización, se utilizó un algoritmo de lematización ligera (*light suffix-stripping*) para eliminar los sufijos morfológicos (en su mayoría inflexiones) como posesivo, plural, relativo, etc. Un segundo lematizador fue utilizado en la colección para eliminar sólo los sufijos plurales. La lista de palabras vacías y el lematizador ligero propuestos están disponibles en la Web[18].

Para indexar los términos, se aplicaron diferentes métodos automáticos de indexación con el objetivo de ser evaluados y comparados. Dos enfoques de indexación

[17] Université.de Neuchâtel, Institut d'informatique - Rue Emile-Argand 11, CH - 2000 Neuchâtel, Suisse

[18] http://www.unine.ch/info/clef/.

independientes del lenguaje que se utilizan son N-gramas (McNamee, et al., 2004) (ver la Sección 3.6.3) y *trunc-n* que es el proceso de truncar una palabra al mantener sus primeros n caracteres y cortar las letras restantes. Se prueban diferentes valores de n, para N-grama y trunc-n, buscando el valor de n que ofrece el mejor rendimiento.

En los experimentos se han implementado seis diferentes modelos de IR para evaluar y comparar los resultados obtenidos. Los modelos son los siguientes:

- El primer modelo es el modelo clásico *idf tf* , donde el peso de cada término indexado t_i es el producto de su frecuencia en el documento D_j (tf_{ij}) por el logaritmo de su frecuencia inversa del documento (idf_j). Se ha normalizado lo ponderación del índice utilizando la normalización del coseno (Manning, et al., 2008).
- Como otro modelo de espacio vectorial, fue adoptado el modelo *Lnu.ltc* sugerido por Singhal (Singhal, 2002). En este modelo se toma en cuenta la longitud del documento. El peso de índice para el término de documento (*Lnu*) se calcula como:

$$w_{ij} = \left[\log\left(tf_{ij}\right) + 1\right].norm_i \qquad (3.9)$$

$$norm_i = \frac{1}{\left(1 + \log\left(\frac{\sum tf_{ij}}{nt_i}\right)\right).((1 - slope).pivot + (slop.nt_i))} \qquad (3.10)$$

donde nt_i es la longitud del documento D_i (número de sus términos de índice), *slope* y *pivot* son constantes. El peso de índice para el término de la consulta (*ltc*) se calcula como:

$$w_{qj} = \left[\log\left(tf_{qj}\right) + 1\right].norm_q\,.idf_j \;;\; norm_q = \frac{1}{\sqrt{\sum_k (tf_{qk}\,.\,idf_j)^2}} \qquad (3.11)$$

- Como el primer modelo probabilístico, fue utilizado el modelo *Okapi (BM25)* sugerido por Robertson (Robertson, et al., 2000). Para este modelo, los parámetros son: b = 0.75, k_1 = 1.2 y *advl* = 202.

- Dos otros modelos probabilísticos, DFR-PL2 y DFR-I(n_e)C2 basados en la medida de la divergencia de aleatoriedad (*DFR, divergence from randomness*) fueron utilizado (Amati, et al., 2002). Aquí tenemos:

$$w_{ij} = Inf_{ij}^1 . Inf_{ij}^2 = -\log_2\left(Prob_{ij}^1(tf_{ij})\right).(1 - Prob_{ij}^2(tf_{ij})) \qquad (3.12)$$

DFR-PL2 se define como:

$$Prob_{ij}^1 = \frac{e^{-\lambda_j}.\lambda_j^{tfn_{ij}}}{tfn_{ij}!} \qquad (3.13)$$

$$Prob_{ij}^2 = \frac{tfn_{ij}}{tfn_{ij}+1} \;\; ; \;\; tfn_{ij} = tf_{ij}\,.\log_2(1 + \frac{c.mean\_dl}{l_i}) \qquad (3.14)$$

y DFR-I(n_e)C2 se define como:

$$Inf_{ij}^1 = tfn_{ij}\,.\log[\frac{n+1}{n_e+0.5}] \;\; ; \;\; n_e = n\,.\left(1 - \left(\frac{n-1}{n}\right)^{tc_j}\right) \qquad (3.15)$$

$$Prob_{ij}^2 = 1 - \frac{1+tc_j}{df_j.(tfn_{ij}+1)} \qquad (3.16)$$

- Por último, se ha empleado uno de los enfoques basados en el modelo del lenguaje (*LM*) sugerido por Hiermstra (Hiemstra, 2000) que se define en la siguiente ecuación (λ_j = 0,35 para todos los términos de índice y *lc* es una estimación de la longitud del corpus C).

$$P(d_i/q) = P(d_i)\,.\prod_{t_j \in q}\left[\lambda_j\,.P(t_j/d_i) + (1-\lambda_j)\,.P(t_j/C)\right] \qquad (3.17)$$

$$P(t_j/d_i) = \frac{tf_{ij}}{l_i} \;\; ; \;\; P(t_j/C) = \frac{df_j}{l_c} \;\; ; \;\; l_c = \sum\nolimits_k df_k$$

Para evaluar el rendimiento de la recuperación se utiliza la medida de precisión promedio (*MAP*) basado en las 100 consultas. El uso de la media proporciona el mismo nivel de importancia para todas las consultas. En la Tabla 3.18 se muestran

los valores de la media de precisión promedio de diferentes modelos con algunos lematizadores (Akasereh, et al., 2012).

Tabla 3.18: Valores de *MAP* para diferentes modelos de IR y lematizadores

Métodos	LM	DFR-PL2	*tf idf*	DFR-I(n_e)C2	Okapi	Lnu-ltc
No lematización/no P.V.	0,3449	0,3905	0,2156	**0,4087**	0,3815	0,3729
No lematización	0,3592	0,4025	0,2648	**0,4069**	0,3962	0,3763
3-gramas	0,3212	0,3743	0,2173	**0,3982**	0,3563	0,3507
4-gramas	0,3325	0,3770	0,2499	**0,4060**	0,3916	0,3574
5-gramas	0,3463	0,3850	0,2581	**0,4068**	0,3911	0,3601
6-gramas	0,3580	0,3963	0,2607	**0,4091**	0,3959	0,3686
Lematizador ligero	0,3668	0,4155	0,2599	**0,4168**	0,4076	0,3874
Lemat. ligero/no P.V.	0,3433	0,3982	0,2040	**0,4117**	0,3785	0,3737
Lematizador de plurales	0,3636	0,4082	0,2696	**0,4124**	0,4010	0,3806
trunc-3	0,3402	0,4000	0,2139	**0,3955**	0,3870	0,3619
trunc-4	0,3635	0,4186	0,2584	**0,4189**	0,4084	0,3862
trunc-5	0,3676	0,4148	0,2687	**0,4185**	0,4077	0,3859
Promedio	0,3506	0,3984	0,2451	**0,4091**	0,3919	0,3718

En referencia a estos resultados, podemos concluir que el mejor modelo de IR es DFR-I(n_e)C2 para cualquier estrategia de lematización y indexación y después los mejores rendimientos globales son, respectivamente, para los modelos *DFR-PL2 y Okapi*.

Otro objetivo del estudio era conocer la diferencia entre los enfoques basados en lematización y los que no incluyen la lematización. La Tabla 3.19 muestra la media del rendimiento de cada estrategia de lematización para los seis diferentes modelos de IR (Akasereh, et al., 2012). Se puede ver también el porcentaje de cambio de este promedio en el rendimiento promedio del caso que no hay lematización.

Tabla 3.19: Rendimiento medio de lematización y su porcentaje de cambio en el caso de no lematización

	Promedio medio de Precisión (*MAP*)							
	LM	**DFR-PL2**	***tf idf***	**DFR-I(n_e)C2**	**Okapi**	**Lnu-ltc**	**Media**	**% cambio no lematizador**
No lema.	0,3592	0,4025	0,2648	0,4069	0,3962	0,3763	0,3677	
3-gramas	0,3212	0,3743	0,2173	0,3982	0,3563	0,3507	0,3363	-8,5%
4-gramas	0,3325	0,3770	0,2499	0,4060	0,3916	0,3574	0,3524	-4,1%
5-gramas	0,3463	0,3850	0,2581	0,4068	0,3911	0,3601	0,3579	-2,7%
6-gramas	0,3580	0,3963	0,2607	0,4091	0,3959	0,3686	0,3648	-0,8%
Lema. ligero	0,3668	0,4155	0,2599	0,4168	0,4076	0,3874	0,3757	+2,2%
Lema. plural	0,3636	0,4082	0,2696	0,4124	0,4010	0,3806	0,3726	+1,3%
trunc-3	0,3402	0,4000	0,2139	0,3955	0,3870	0,3619	0,3498	-4,9%
trunc-4	0,3635	0,4186	0,2584	0,4189	0,4084	0,3862	0,3757	+2,2%
trunc-5	0,3676	0,4148	0,2687	0,4185	0,4077	0,3859	**0,3772**	**+2,6%**

Basándose en estos valores, la estrategia trunc-5 tiene el mejor promedio de 0,3772 y se mejora de 2,6% con respecto a otros modelos. Con una pequeña diferencia trunc-4 y el lematizador ligero son los próximos que se mejoran de 2,2%. Considerando el modelo de mejor rendimiento (DFR-I (n_e)C2), sus resultados de promedio medio de precisión con diferentes estrategias y sus porcentajes de cambio en el caso de no lematización muestran, de nuevo, que las estrategias trunc-4 y trunc-5 son los que más mejoran con respecto al caso de no lematización y trunc-3 y de 3-gramas disminuyen el rendimiento en comparación con el caso de no lematización. Otras datos de estos experimentos muestran que para todos los modelos excepto DFR-I (n_e) C2 y *tfidf* , los peores resultados son cuando se aplica el método 3-gramas. Para DFR-I (n_e) C2 y modelos *idftf* el peor rendimiento se resulta de trunc-3.

Con el fin de analizar el efecto de la eliminación de palabras vacías en la eficacia de la recuperación, se aplicó la estrategia de no lematización y lematización ligera a los seis modelos de IR con y sin eliminación de palabras vacías. La Tabla 3.20

indica los resultados de prueba sin lematizador y en la Tabla 3.21 se muestra los resultados de experimentos con uso de lematizador ligero.

Tabla 3.20: Valores de *MAP* en el caso de no lematización con y sin eliminación de palabras vacías

	Promedio medio de Precisión (*MAP*)						
	LM	**DFR-PL2**	***tf idf***	**DFR-I(n_e)C2**	**Okapi**	**Lnu-ltc**	**Media**
No lemat./no P.V.	0,3449	0,3905	0,2156	**0,4087**	0,3815	0,3729	0,3524
No lemat. + P.V.	0,3592	0,4025	0,2648	**0,4069**	0,3962	0,3763	0,3677
% de cambio	+4,1%	+3,1%	+22,8	-0,4%	+3,9%	+0,9%	+4,3%

Tabla 3.21: Valores de *MAP* en el caso de lematización ligera con y sin eliminación de palabras vacías

	Promedio medio de Precisión (*MAP*)						
	LM	**DFR-PL2**	***tf idf***	**DFR-I(n_e)C2**	**Okapi**	**Lnu-ltc**	**Media**
Lemat. ligero/no P.V.	0,3433	0,3982	0,2040	**0,4117**	0,3785	0,3737	0,3516
Lemat. ligero + P.V.	0,3668	0,4155	0,2599	**0,4168**	0,4076	0,3874	0,3757
% de cambio	+6,8%	+4,3%	+27,4%	+1,2%	+7,7%	+3,7%	+6,9%

Obviamente, la eliminación de palabras vacías ayuda a mejorar la eficacia de la recuperación, ya que aplicando la eliminación de palabras vacías se produce un aumento del rendimiento promedio del 4,3% (sin lematización) y del 6,9% (con lematización ligera). Los resultados revelan que para ambos enfoques sin eliminación de palabras vacías, el modelo DFR-I (n_e)C2 tiene todavía el valor más alto de *MAP* en comparación con otros modelos. De hecho, la aplicación de la eliminación de palabras vacías tiene un pequeño impacto en el *MAP* para este modelo (-0,4% para no lematización y 1,2% para la lematización ligera). La eliminación de palabras vacías tiene su mayor impacto en el modelo *idf tf*. En este modelo, como la ponderación de término depende de la frecuencia de los términos, obviamente, en un texto sin ruido de las palabras vacías se puede realizar el cálculo más exacto de peso de los términos y similitud. Otra conclusión que se puede alcanzar a partir de los resultados obtenidos es

que la aplicación de la eliminación de palabras vacías sin realizar ninguna lematización produce un mejor resultado que la aplicación de lematizador ligero sin la eliminación de palabras vacías. La razón es que muchos sufijos no adjuntados a las palabras que se incluyen en la lista de palabras vacías y así desaparecen por la eliminación de palabras vacías y no por la lematización ligera.

Observaciones

En estos experimentos se utilizan dos lematizadores uno que elimina sólo los signos plurales y otro quita los sufijos. Aunque ambos no son completos, es decir, el primero no considera todos los signos plurales y el otro es un lematizador ligero, consiguen mejorar el rendimiento de recuperación en el texto persa. No ha ninguna explicación de cómo se identifican las palabras vacías pero parece que son las palabras más frecuentes en el corpus. Suponiendo que las palabras vacías son las más frecuentes en la colección entonces otra conclusión que podemos tener a partir de los resultados indicados en las tablas 3.20 y 3.21 es que la tokenización del texto no es tan correcta, porque los sufijos están considerados separados de sus palabras y eso aumenta el número de recurrencia de los sufijos y desde luego se ponen en la lista de palabras vacías. Sin embargo los resultados revelan que, para la lengua persa, la eliminación de palabras vacías desempeña un papel importante en la mejora del rendimiento de la recuperación de información.

En cuanto a los modelos de RI utilizados en estos experimentos notamos que es simplemente la aplicación de los modelos, sin ninguna modificación particular de adaptación al persa. La comparación de estos modelos muestra que, por lo general, los modelos probabilísticos basados en paradigma DFR (*Divergence From Randomness*) tienen mejores resultados de recuperación para cualquier estrategia de indexación y lematización. El modelo DFR-I (n_e)C2 fue el mejor modelo seguido del modelo DFR-PL2.

3.6.5. Detección de Documentos Similares en la RI en Persa

Otro trabajo de investigación en el campo de RI realizado por la universidad de ciencias y tecnologías de Irán es cómo optimizar la detección de documentos similares en la RI para documentos persas (Kashefi, et al., 2010). En realidad, este trabajo consiste en evaluar la medida de similitud entre los documentos antes y después de la eliminación de afijos en el texto persa. La evaluación de la eficacia de la eliminación de afijos se hace utilizando cuatro enfoques principales de la similitud de documentos; indexación semántica latente, modelo de espacio vectorial, *Shingling* y co-ocurrencia.

En el caso del modelo de espacio vectorial, el documento se representa como un vector de las palabras y se calcula la frecuencia de cada palabra en los documento de la colección. Si las palabras son elegidas como términos, entonces cada palabra en el vocabulario se convierte en una dimensión independiente en un espacio vectorial dimensional muy alto. La similitud entre dos documentos se determina por una medida de similitud como el valor de coseno entre sus correspondientes vectores.

La Indexación Semántica Latente (*LSI*) fue descrita originalmente por Deerwester et al. (Deerwester, et al., 1990). En los modelos clásicos de RI la relevancia de un documento a una consulta está basada por el número de concurrencias de palabras de la consulta que se encuentren en los documentos. Un defecto de esta forma de medir la relevancia es que no se tiene en cuenta el contexto semántico de la palabra, y como consecuencia de esto van a aparecer dos problemas en la búsqueda de información, la sinonimia (términos distintos con el mismo significado) y la polisemia (términos iguales con distintos significados). La indexación semántica latente, propone un método para solucionar estos problemas. La idea es pasar de un conjunto de términos a un conjunto de entidades donde podamos sacar la estructura latente en la asociación entre términos y documentos.

W-Shingling es otro método de similitud de documentos que permite definir una medida de similitud y de inclusión entre documentos (Broder, et al., 1997). Para ello, cada documento es convertido en una secuencia canónica de *tokens*. Esta secuencia sólo contiene las palabras del documento. A partir de esta secuencia de

tokens, se define el concepto de *shingle* que es una sub-secuencia continua de w palabras. Luego se define el conjunto *w-shingling* S(D,w) como el conjunto de todos los *shingles* de longitud w. Por ejemplo dada la secuencia (hola, que, tal) el conjunto 2-*shingle* sería {(hola, que), (que, tal)}.

Dados dos documentos *A* y *B* se define la similitud (r de "*resemblance*") entre ellos como:

$$r(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|} \qquad (3.18)$$

Otro enfoque básico para el cálculo de la similitud de documentos es el método de "co-ocurrencia" que ya está aplicado para la lengua persa por el trabajo de Zamanifar (Zamanifar, et al., 2008). Este método tiene tres pasos principales: la identificación de tema, la interpretación de tema y la medida de similitud. Identificación de temas comprende la segmentación del texto, la eliminación de palabras triviales o redundantes y la lematización. La parte principal del método es la interpretación de tema que consiste en la búsqueda y la clasificación de palabras importantes. Al final, el grado de similitud entre las frases se calcula sobre la base de medida de similitud. Este método está construido sobre la cadena léxica de palabras importantes y está basado en la propiedad de co-ocurrencia de términos. Se evita que los documentos irrelevantes sean identificados similares debido a la propiedad polisemia de las palabras. Considera también el orden de las palabras en la identificación de los documentos similares.

A través de un estudio exhaustivo sobre los afijos y sus diferentes composiciones en la lengua persa, se extrajeron 300 afijos flexivos comunes para construir un algoritmo de eliminación de afijos. En primero, para evaluar el algoritmo propuesto, los afijos de 10 pares de documentos similares fueron editados por un experto y por la aplicación del método de extracción de afijos. La Tabla 3.22 muestra el número de afijos eliminados en ambos casos (Kashefi, et al., 2010).

Tabla 3.22: Comparación de afijos eliminados

Documentos	N°. de afijos por el experto	Nº. de afijos por el algoritmo
Doc. 1	25	28
Doc. 2	27	31
Doc. 3	16	15
Doc. 4	24	24
Doc. 5	29	30
Doc. 6	26	27
Doc. 7	31	31
Doc. 8	24	24
Doc. 9	19	20
Doc. 10	25	23

La segunda evaluación fue la medida de similitud entre 50 pares de documentos similares antes y después de la extracción de afijos, utilizando los 4 métodos mencionados anteriormente. La evaluación se basa simplemente en el cálculo de la diferencia de los valores de similitud entre los documentos antes de eliminar los sufijos y después de eliminarlos. La Tabla 3.23 compara los resultados promedios de cuatro métodos (Kashefi, et al., 2010). Los resultados son los valores de similitudes normalizadas es decir, el valor 1 significa exactamente lo mismo documento y el valor 0 significa no similar.

Observaciones

En un sistema de RI para identificar los documentos similares existen muchas técnicas que fueron desarrollados a lo largo del tiempo. Unos de los métodos, como los métodos descritos en este trabajo, están basados en palabras (*word-base*) que se enfocan en la co-ocurrencia de los términos en los textos. La eliminación de los afijos es la parte de la lematización que reduce las palabras flexivas (o a veces derivadas) a su lema, forma base o raíz. En efecto, la aplicación de este proceso sobre los documentos de una colección permite tener palabras parecidas en diferentes documentos y por secuencia podemos detectar los documentos similares. En realidad, este trabajo es sólo la

comparación de cuatro modelos de recuperación con y sin aplicación de lematización sobre los documentos persas. Los resultados obtenidos (ver la Tabla 3.23) muestran que la eliminación de afijos en los documentos mejora la detección y recuperación de documentos similares para todos estos cuatro métodos. Parece que esta mejora no es muy importante porque las 300 reglas de afijos son bastante pocas para poder generar todas las reglas gramaticales y excepciones de la lengua persa en hacer algoritmos de lematización.

Tabla 3.23 Comparación de similitud entre documentos similares antes y después de eliminar los afijos

Método de similitud	Antes de eliminar los afijos	Después de eliminar los afijos
ISL	0,700	0,803
Modelo de espacio vectorial	0,728	0,810
W-Shingling	0,268	0,312
Co-ocurrencia	0,781	0,871

3.7. Conclusiones

Este capítulo se ha realizado un análisis de los trabajos previos realizados en el ámbito de la recuperación de información para la lengua persa. Como ha quedado patente, hay relativamente pocos estudios sobre la recuperación de los documentos persas y una de las principales razones es la falta de conjuntos de prueba estándar. La primera necesidad de los investigadores en el campo de IR para poder desarrollar y probar las herramientas es tener o poder acceder a colecciones estándares de prueba. La elección de una colección adecuada es de la suma importancia a la hora de evaluar un sistema, ya que únicamente así tendremos la convicción de que los resultados obtenidos son fiables y representativos. La mayoría de las colecciones de prueba en persa son pequeñas colecciones que fueron construidas en cada centro de investigaciones para poder probar sus propias herramientas desarrolladas y en algunos casos no están disponibles al público. Además, los documentos de estas colecciones fueron recogidos manualmente sin verificar la calidad y equilibrio de ellos.

Entre todas las colecciones de prueba destaca el corpus Hamshahri que es una colección construida, hace poco tiempo, según las especificaciones de TREC y desde luego se utiliza como una fuente de documentos persas para los investigadores de RI. La versión original fue construida en 2007 y los documentos eran en formato texto. Esta versión tiene más de 160.000 documentos y 65 temas. Otra versión (versión 1 de *CLEF2008*) fue preparada en el año 2008 que contiene 100 temas, el juicio de relevancia y el mismo número de documentos que la versión original pero en formato XML. La segunda versión (versión 2 de *CLEF2009*) fue preparada en el año 2009 y contiene 50 temas, el juicio de relevancia y alrededor de 320.000 documentos XML. Los documentos de esta última versión contienen imágenes.

El primer paso realizado en un proceso de RI es la tarea de tokenización que consiste en la separación de palabra, identificación de *tokens* y tipos de palabras. La lengua persa debido a sus morfologías complejas, diferentes formas de letras, uso o eliminación de los espacios y el uso de diversas formas de caracteres en una palabra, tiene muchas dificultades en la identificación de las palabras en un texto. Hay sólo dos trabajos (Megerdoomian, et al., 2000) y (Shamsfard, et al., 2010) dedicados a la tokenización del texto persa pero como estos tokenizadores no son disponibles para público, resulta difícil de hacer una evaluación comparativa de ellos. Lo que es evidente es que el texto persa necesita una pre-normalización para llegar a un texto estándar de tal manera que el proceso de tokenización sea más fácil. La pre-normalización del texto se hace mediante herramientas que consideran las características de la lengua y que, en el caso de la lengua persa, necesita una base de datos para manejar las excepciones y la diversidad de la forma escritura o diferentes estilos de ortografía de palabras.

La identificación de palabras vacías es otra de las tareas principales en la construcción de un sistema de RI. Las listas de palabras vacías persas disponibles son casi todas deducidas de pequeñas colección de textos basándose en la frecuencia de los términos en la colección. Hay algunas que son muy cortas e incompletas y otras que fueron construidas manualmente considerando el significado de la palabra. Sin embargo, el uso de una única lista de palabras vacías a través de diferentes colecciones de documentos persas podría ser perjudicial para la eficacia de recuperación en un sistema de RI. Por le tanto, es preferible poder derivar una lista de palabras vacías para

una determinada colección. No hay ningún trabajo de investigación que consista en identificar automáticamente las palabras vacías persas en un sistema de RI. Por eso se propone en al capítulo 5 un método automático para identificar las palabras vacías para sistemas de RI en persa.

Respecto a la lematización del texto persa, hay diferentes algoritmos donde la mayoría de los lematizadores son basados en reglas similares al algoritmo de Porter. Este enfoque de algoritmo quita el prefijo o sufijo de acuerdo con un conjunto de reglas gramaticales. Hay unos algoritmos que son lematizadores ligeros que tratan sólo los afijos y unos signos plurales. Hay otros que son más completos que tratan también los verbos y utilizan una base de datos para las excepciones. La mayoría de estos lematizadores fueron evaluados usando un subconjunto limitado de palabras o un corpus limitado y los métodos de evaluación eran diferentes. Desgraciadamente, no tenemos acceso a las fuentes de todos estos algoritmos para poder comprobarlos en la misma colección de documentos. La eficacia de un algoritmo de lematización se debe evaluar en un sistema de IR utilizando las medidas estándares (como la precisión y la exhaustividad).

En cuanto a los modelos de la RI, los modelos aplicados a los documentos persas son los mismos modelos que fueron creados por los investigadores a lo largo del tiempo. Los modelos fueron aplicados en diferentes sistemas de RI con documentos persas. Cada sistema tenía su propia colección de documentos, sus consultas y sus métodos de evaluación. Por lo tanto, no es razonable juzgar esos modelos ya que no se pueden probarlos en el mismo sistema de RI. Pero de todos estos experimentos tenemos la fuerte convicción que ambas operaciones, la eliminación de palabras vacías y la lematización del texto ayudan a mejorar la eficacia de recuperación para cualquier modelo para documentos escritos en persa.

De los modelos que fueron aplicados sobre los documentos de un corpus estándar, es decir la colección Hamshahri, podemos tener unas conclusiones que son las siguientes:

- Considerando la precisión de los 20 primeros documentos recuperados sin tener en cuenta del valor de exhaustividad, el rendimiento del modelo difuso es mejor que el modelo de espacio vectorial aplicando a los documentos persas.
- Basándose siempre sobre la precisión de los 20 primeros documentos recuperados, podemos decir que el modelo de análisis del contexto local mejora marginalmente en comparación con el modelo de espacio vectorial. Esto podría ser debido al hecho de que el método de ponderación *Lnu.ltu* está funcionando bien en el idioma persa. Otra explicación también podría ser que los parámetros utilizando por el método de *LCA* son los mismos que se utilizan en *TREC* y puede ser que necesite una modificación de los parámetros para la colección de Hamshahri.
- Otra conclusión es que el modelo de 4-gramas basado en el espacio vectorial con esquema de ponderación *Lnu.ltu* tiene mejores resultados que 3-gramas incluso la configuración basada en los términos y que esquema de ponderación *Lnu.ltu* tienen considerablemente mejor rendimiento que *atc.atc*.
- La aplicación de métodos tales como N-gramas (especialmente con un pequeño valor de n) disminuyen claramente el rendimiento de la recuperación en comparación con el caso que no haya la eliminación de palabras vacías y tampoco la aplicación de ningún lematizador. Eso se puede explicar por el hecho de que las palabras vacías persa tienen muy pocas letras (ver la Sección 5.3.1) y la aplicación de N-gramas (N de pequeño valor) o trunc-3 será generalmente sobre las palabras vacías que no son palabras significativas. Otra explicación es que la lematización permite disminuir la presencia en el texto de palabras con muy pocas letras. Estas palabras pueden ser los sufijos como por ejemplo los signos plurales que están separados de las palabras originales en la fase de tokenización del texto.

Capítulo 4

Documentos Persas y los Buscadores Web

Resumen

La lengua persa es diferente de las lenguas occidentales especialmente en las variaciones de la morfología y de la ortografía. De hecho, el rendimiento de los sistemas de recuperación de información en el idioma persa sigue siendo problemático. Por esta razón, estamos interesados en estudiar la evaluación de la recuperación de información de los documentos persas en la Web. El motor de búsqueda Google, como una herramienta de recuperación de información en la Web, es utilizado por casi el 92% de los usuarios iraníes. Entonces, el objetivo principal de este capítulo es analizar la calidad de los buscadores Web y especialmente el buscador de Google en la búsqueda de información en persa. La determinación del rendimiento de la recuperación se basó en el cálculo de precisión y exhaustividad relativas de los resultados de búsquedas realizadas en un sitio web que hemos construido con los documentos del corpus Hamshahri. Preguntamos al buscador Google 100 consultas del corpus y comparamos las páginas web devueltas con los documentos relevantes del juicio de relevancia del corpus. Los resultados obtenidos muestran que Google no toma en cuenta correctamente el análisis morfológico de la lengua persa. El error en la segmentación de palabras del texto, la consideración de palabras vacías como palabras claves del contenido de un documento y las variantes erróneas de palabras claves encontradas por Google son las principales razones que afectan al rendimiento de la recuperación de información en persa.

4.1. Introducción

En los últimos tiempos, Internet se ha convertido en una de las principales fuentes de información de nuestra sociedad, albergando en sus servidores millones de documentos por lo que se ha convertido en una odisea recuperar y organizar la información existente. Se ha estimado que hay aproximadamente entre 4,5 y 5 billón páginas presentes en la Web (World Wide Web Size, 2015). De acuerdo con los estudios, sólo el 80-85% de las páginas web que están disponibles en la Web dan información útil, y

el 20-15% restantes son en su mayoría los duplicados de las páginas originales o cerca de duplicados, mientras que algunos de ellos son páginas irrelevantes completamente (Garg, et al., 2012). Por lo tanto, la explosión de Internet ofrece una gran cantidad de nuevos problemas para los sistemas de recuperación de información. Estos sistemas de recuperación de información ayudan a los usuarios completar las tareas de búsqueda, mediante la búsqueda de un puñado de los documentos pertinentes, entre miles y miles de páginas de texto con poca organización estructural. Al mismo tiempo, los desarrolladores de sistemas de recuperación deben ser capaces de evaluar la eficacia general de estos sistemas es decir, la relevancia de los resultados se recupera en respuesta a una consulta del usuario.

4.2. La Recuperación de Información en la Web

La búsqueda de información en la Web es una práctica común para los usuarios de Internet y los sistemas de recuperación de información Web (conocidos como motores de búsqueda) se han convertido en herramientas indispensables para los usuarios. Su arquitectura y modo de operación se basan en poder recolectar mediante un mecanismo adecuado los documentos existentes en los sitios web. Una vez obtenidos, se llevan a cabo tareas de procesamiento que permiten extraer términos significativos contenidos dentro de los mismos, junto con otra información, a los efectos de construir estructuras de datos (índices) que permitan realizar búsquedas de manera eficiente. Luego, a partir de una consulta realizada por un usuario, un motor de búsqueda extraerá de los índices las referencias que satisfagan la consulta y se retornará una respuesta ordenada por diversos criterios al usuario.

Sin embargo, la RI en la Web ha sido siempre tarea diferente y difícil en comparación con un sistema de RI clásica. Baeza-Yates (Baeza-Yates, et al., 1999) plantea que hay desafíos de dos tipos. Básicamente las diferencias se pueden dividir en dos partes, a saber, las diferencias en los documentos y las diferencias en los usuarios.

a) Respecto de los datos

Distribuidos: La Web es un sistema distribuido, donde cada proveedor de información pública su información en computadoras pertenecientes a redes conectadas a Internet, sin una estructura ó topología predefinida.

Volátiles: La dinámica del sistema hace que exista información nueva a cada momento o bien que cambie su contenido o inclusive desaparezca otra que se encontraba disponible.

No estructurados y redundantes: Básicamente, la Web está formada de páginas HTML, las cuales no cuentan con una estructura única ni fija. Además, más del 20% de los documentos presentes son duplicados y esta estimación no incluye los duplicados semánticos (Garg, et al., 2012).

Calidad: En general, la calidad de la información publicada en la Web es altamente variable, tanto en escritura como en actualización (existe información que puede considerarse obsoleta), e inclusive existe información con errores sintácticos, ortográficos y otros.

Heterogeneidad del documento: El contenido de una página Web es de naturaleza heterogénea, es decir, además de texto puede contener otros contenidos multimedios como audios, vídeos e imágenes.

Hipertexto: Los documentos presentes en la Web son diferentes de los documentos habituales que son textos, debido a la presencia de hiperenlaces. Se estima que hay aproximadamente una media de 10 hipervínculos por documento (Garg, et al., 2012).

Número de documentos: El tamaño de la Web ha crecido exponencialmente en los últimos años. El número de documentos es más de billón y esta colección es mucho mayor que cualquier colección de documentos procesados por un sistema de RI. De acuerdo con la estimación, Web crece actualmente en un 10% por mes (Garg, et al., 2012).

b) Respecto de los usuarios

Los usuarios en la Web se comportan de manera diferente que los usuarios de los sistemas de RI clásicos. Los usuarios de estos últimos son en su mayoría bibliotecarios mientras que la gama de usuarios de la Web varía de un principiante a una persona técnicamente sólida.

Especificación de la consulta: Los usuarios encuentran dificultades para precisar, en el lenguaje de consulta, su necesidad de información. La mayor parte de las consultas enviadas por los usuarios son generalmente cortas y carecen de palabras claves útiles que pueden ayudar en la recuperación de información relevante.

Manejo de las respuestas: Cuando un usuario realiza una consulta se ve sobrecargado de respuestas, siendo una parte de las mismas irrelevante. Por lo general, los usuarios no evalúan todas las pantallas de resultados, restringen sólo a los resultados que se muestran en la primera pantalla de resultados.

Heterogeneidad de los usuarios: Hay una amplia variación en la educación y la experiencia de web entre los usuarios de la Web.

Por lo tanto, el principal reto de la recuperación de información en la Web es la forma de satisfacer las necesidades de los usuarios, dada la heterogeneidad de las páginas web y las consultas de baja calidad.

4.3. Herramientas de la Recuperación de Información en la Web

La información en la Web puede ser recuperada por un variado conjunto de herramientas disponibles, que van desde los motores de búsqueda de propósito general a motores de búsqueda especializados. Las más utilizadas son los motores de búsqueda de uso general (*General-Purpose Search Engine*) como Google, Altavista, Excite, Yahoo, Bing. Cada uno de ellos tiene su propia página web que permite realizar búsquedas a una consulta en la red.

Un motor de búsqueda es un sistema informático que busca archivos almacenados en servidores web, por lo que es posible encontrar un conjunto de los recursos que responden a una consulta del usuario (Hajjar, et al., 2014). Estos recursos pueden ser páginas web, imágenes, videos, archivos, etc., que están representados por documentos de diferentes formatos (HTML, JPEG, MPEG, PDF, etc.). La importancia de este motor depende de la relevancia del resultado global que puede contener millones de páginas web. Algunas páginas pueden ser más pertinentes y accesibles que otras.

Los motores de búsqueda o *search engines* se basan en un robot o software que recorre la red automáticamente para localizar documentos, los indiza y los introduce en una base de datos. Esta base de datos será interrogada por los usuarios a través de un formulario o interfaz web, que lanza la búsqueda, la compara con los recursos indizados en la base de datos y devuelve como resultado un conjunto de enlaces. La forma de búsqueda en estos sistemas es a través de palabras clave introducidas en el formulario de consulta, permitiendo la mayoría de ellos realizar búsquedas simples y avanzadas.

4.3.1. Funcionamiento de los Motores de Búsqueda

Un motor de búsqueda en Internet se compone de diversos elementos, los cuales serán evaluables, a la hora de valorar su rendimiento o utilidad a la hora de satisfacer una demanda de información (ver la Figura 4.1):

- El **robot** que recorre Internet para localizar direcciones y documentos y que genera una base de datos textual.
- Un **sistema de indexación automática**, según distintos criterios (texto completo, parcial o utilizando las etiquetas propias del lenguaje de marcas).
- Un **motor de búsqueda** o *search engine.*
- Un **sistema de interrogación** que incluye un lenguaje de consulta y una serie de procedimientos más o menos documentales para precisarlas.

- Un **interfaz**, permitiendo al usuario hacer peticiones de búsqueda y presentar los resultados.

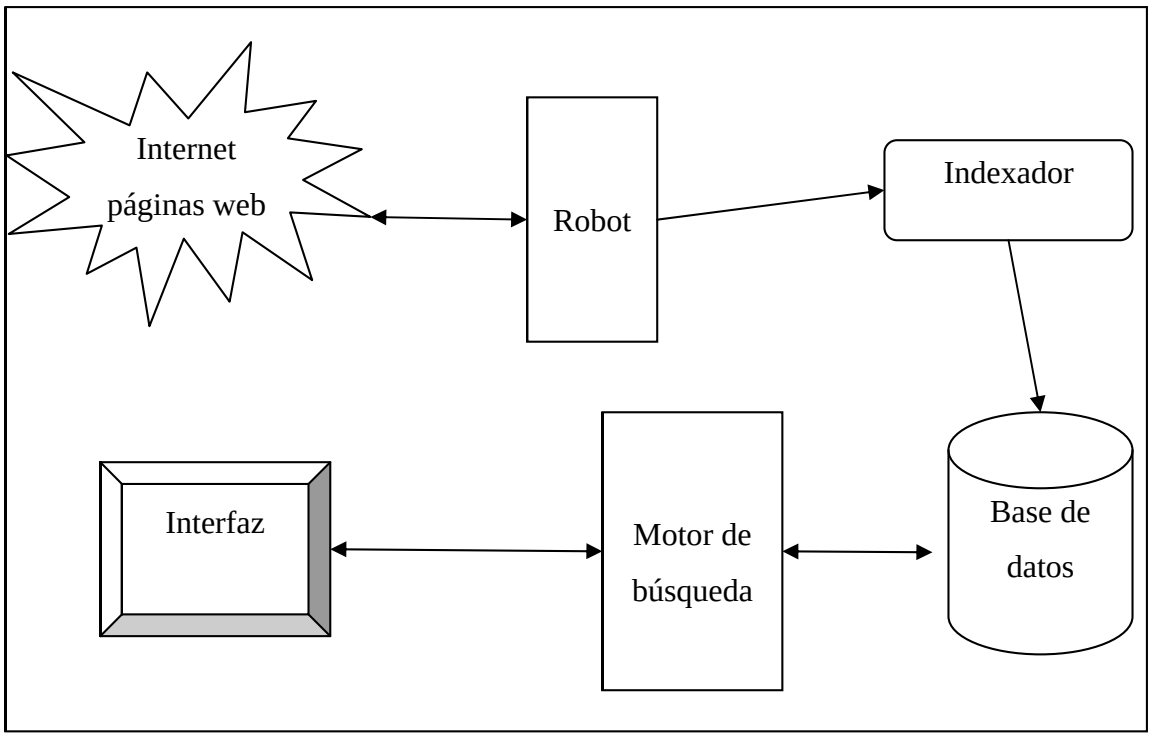

Figura 4.1: La arquitectura de un buscador de web

4.4. Internet en Irán

En 1993, Irán se convirtió el segundo país del oriente medio de estar conectado a Internet y desde entonces el gobierno ha realizado importantes esfuerzos para mejorar la infraestructura de la telecomunicación nacional (Wikipedia, 2015). La infraestructura nacional de la conexión a Internet se basa en dos grandes redes: la red telefónica pública conmutada *PSTN* (*Public Switched Telephone Network*) y la red de datos públicos. La red *PSTN* proporciona una conexión para los usuarios finales a los proveedores de servicios de Internet (*ISP, Internet Service Provider*) a través de líneas digitales y conexiones basadas en módem. La empresa de comunicación de datos de Irán *DCI* (*Data Communication Company of Iran*), una subsidiaria de la compañía de comunicación de Irán *TCI* (*Telecommunication Company of Iran*), opera la red de datos públicos.

Tabla 4.1: Crecimiento de Internet en Irán

Año	N° de usuarios	Población	Penetración (% población)	Fuente de uso
2000	250.000	69.442.905	0,36 %	ITU[19]
2002	5.500.000	69.442.905	7,5 %	ITU
2005	7.500.000	69.442.905	10,8 %	ITU
2008	23.000.000	65.875.223	34,9 %	ITU
2009	32.200.000	66.429.284	48,5 %	IWS[20]
2010	33.200.000	76.923.300	43,2 %	IWS
2012	42.000.000	78.868.711	53,3 %	IWS
2014	45.000.000	80.840.713	55,7 %	IWS

El primer uso público de Internet en el país comenzó en 1995 en las universidades a través de acceso telefónico. La demanda de Internet aumentó rápidamente, haciendo de Internet muy popular en pocos años. Los datos de la Tabla 4.1 nos muestran la estadística de la población y el crecimiento de Internet en la sociedad iraní. Los usuarios de Internet han aumentado desde el año 2000 y como se ha indicado en la Sección 1.2 según *Internet World Stats* (ver la Tabla 4.2), Irán es actualmente el 13º país en el mundo con el mayor número de usuarios de Internet (Internet World Stats, 2013).

[19] ITU (International Telecommunication Union) es el organismo especializado de las Naciones Unidas para las tecnologías de información y comunicación.

[20] Internet World Stats

Tabla 4.2: Los 20 primeros países con mayor número de usuario de Internet, datos de 31/12/2013

#	País o región	Población 2013 estimación	Usuarios Internet año 2000	Usuarios Internet últimos datos	Penetración (%población)	Usuario % mundo
1	China	1.343.239.923	22.500.000	538.000.000	40,1 %	22,4 %
2	Estados unidos	313.847.465	95.354.000	245.203.319	78,1 %	10,2 %
3	India	1.205.073.612	5.000.000	137.000.000	11,4 %	5,7 %
4	Japón	127.368.088	47.080.000	101.228.736	79,5 %	4,2 %
5	Brasil	193.946.886	5.000.000	88.494.756	45,6 %	3,7 %
6	Rusia	142.517.670	3.100.000	67.982.547	47,7 %	2,8 %
7	Alemania	81.305.856	24.000.000	67.483.860	83,0 %	2,8 %
8	Indonesia	248.645.008	2.000.000	55.000.000	22,1 %	2,3 %
9	Reino unido	63.047.162	15.400.000	52.731.209	83,6 %	2,2 %
10	Francia	65.630.692	8.500.000	52.228.905	79,6 %	2,2 %
11	Nigeria	170.123.740	200.000	48.366.179	28,4 %	2,0 %
12	México	114.975.406	2.712.400	42.000.000	36,5 %	1,7 %
13	**Irán**	**78.868.711**	**250.000**	**42.000.000**	**53,3 %**	**1,7 %**
14	Corea	48.860.500	19.040.000	40.329.660	82,5 %	1,7 %
15	Turquía	79.749.461	2.000.000	36.455.000	45,7 %	1,5 %
16	Italia	61.261.254	13.200.000	35.800.000	58,4 %	1,5 %
17	Filipinas	103.775.002	2.000.000	33.600.000	32,4 %	1,4 %
18	España	47.042.984	5.387.800	31.606.233	67,2 %	1,3 %
19	Vietnam	91.519.289	200.000	31.034.900	33,9 %	1,3 %
20	Egipto	83.688.164	450.000	29.809.724	35,6 %	1,2 %

4.5. Presencia de la Lengua Persa en la Web

La lengua persa, ahora, es la lengua de muchos documentos publicados en la red y, aproximadamente, se utiliza por el 0,8% de todos los sitios web (W3Techs, 2015). La Tabla 4.3 muestra el porcentaje de los sitios web que utilizan varios lenguajes de contenido. El número de blogs en persa también ha experimentado un crecimiento espectacular, elevando el persa a ser uno de los diez idiomas de la blogosfera mundial (Megerdoomian, 2008).

Tabla 4.3: Uso de las lenguas de contenido para sitios web, datos del octubre 2014

Idioma	Porcentaje de los sitios web
Inglés	55,7
Alemán	6,1
Ruso	5,7
Japonés	5,1
Español, castellano	4,8
Francés	4,1
Chino	2,8
Portugués	2,4
Italiano	1,9
Polaco	1,7
Turco	1,4
Holandés, flamenco	1,3
Persa	**0,8**
Árabe	0,8
Checo	0,7
Coreano	0,5
Sueco	0,5
Indonesio	0,4
Vietnamita	0,4
Rumano	0,4

4.6. Recuperación de los Documentos Persas en la Web

Gran parte de las necesidades de recuperación de información para los usuarios está resuelta con la implementación de los buscadores de web. El rendimiento de los motores de búsqueda varía con el lenguaje utilizado, y depende de la naturaleza y la complejidad de la lengua en la que se formula la solicitud de búsqueda (Hajjar, et al., 2014). La operación de un motor se basa principalmente en un tratamiento automático del lenguaje natural. Estos tratamientos difieren de un idioma a otro, y pueden depender de las características particulares de este idioma (Wikipedia, 2013). Así que es fácil ver el papel de la estructura de un lenguaje natural en el acceso a la información en el

documento. El rendimiento de los motores de búsqueda depende principalmente de la eficacia de los métodos de indexación y la recuperación de la información, que constituyen el núcleo de estos sistemas (Hajjar, et al., 2014). La mayoría de los motores de búsqueda disponibles que se desarrollan principalmente para los idiomas occidentales tales como el inglés, son más potentes analizar los documentos escritos en estos idiomas. Además, estas características son peores en el caso de la lengua persa, probablemente debido a las diferencias de las especificidades morfológicas y características estructurales de persa en comparación con las lenguas occidentales. De hecho, en esta parte de nuestro trabajo, estamos interesados hacer un estudio sobre la evaluación de la recuperación de información en persa a través de un análisis del rendimiento del motor de búsqueda en la extracción de información relevante sobre documentos en persa.

4.6.1. El Motor de Búsqueda Google

Desde las estadísticas hechas en el mes de agosto de 2010 por *StatCounter Global Stats*[21], 91,56% de los usuarios en Irán utilizan Google seguido de Yahoo (4,28%) y Bing (3,53%). Según otro estudio realizado por Tawileh (Tawileh, et al., 2010) sobre la evaluación de los motores de búsqueda para el idioma árabe revela que, casi todas las veces, los resultados de Google son mejores que otros buscadores. Puesto que la escritura del idioma persa se parece más a la lengua árabe y hay bastante palabras árabes que se utilizan en el texto persa, entonces estas dos conclusiones pueden ser motivos suficientes para justificar nuestro estudio de la eficiencia de Google con respecto a la búsqueda de documentos persas en la Web.

Google es uno de los motores de búsqueda en Internet más grande, más rápidos y más utilizados en la actualidad. Ofrece una forma rápida y sencilla de encontrar información en la Web. Según las estadísticas del mes de agosto de 2012,

[21] http://ptgmedia.pearsoncmg.com/images/9780789747884/supplements/9780789747884_appC.pdf

http://gs.statcounter.com/

Google tiene acceso a un índice de 30 billón de páginas web y responde a más de 100 mil millones de consultas al mes (3,3 mil millones de búsquedas por día y más de 38.000 mil por segundo) (Sullivan, 2012). Tiene una forma muy particular para establecer la relevancia de los resultados: utiliza el número de enlaces de una página concreta como medida para evaluar su calidad informativa (Brin, et al., 1998). De este modo, cada vínculo de una página a otra funciona como un voto a favor de la página receptora. Además el Google no valora todos los votos por igual: valen más aquellos vínculos, o votos, que provengan de páginas que a su vez reciban más enlaces de otras páginas. La popularidad de Google (disponible en muchos idiomas, entre ellos persa) se ha extendido por la red en un tiempo récord.

4.7. Evaluación del Buscador Google en Documentos Persas

Hay dos aspectos principales para medir el rendimiento del sistema de RI: eficiencia y eficacia. La eficiencia puede ser medida en términos de tiempo (por ejemplo, segundo por consulta) y el espacio (por ejemplo, bytes por documento). El aspecto más visible de la eficiencia es el tiempo de respuesta (también conocido como latencia) experimentada por un usuario entre la emisión de una consulta y la recepción de los resultados. Cuando muchos usuarios simultáneos deben ser atendidos el rendimiento de consulta, medido en consultas por segundo, se convierte en un factor importante en el rendimiento del sistema. Para un motor de búsqueda web de propósito general, el rendimiento requerido puede ir mucho más allá de las decenas de miles de consultas por segundo. La eficiencia también puede considerarse en términos del espacio de almacenamiento, medida por los bytes de disco y la memoria requerida para almacenar el índice y otras estructuras de datos. Además, cuando miles de máquinas están trabajando en conjunto para generar un resultado de búsqueda, su consumo de energía y la huella de carbono también se convierten en consideraciones importantes.

En nuestro caso, el más importante es saber si un documento devuelvo por Google es relevante a una dada consulta. La idea clave detrás de la medida de la efectividad es la noción de relevancia. El aspecto de relevancia se mide por la eficacia.

La eficacia es más difícil de medir que la eficiencia ya que depende enteramente de juicio humano. Un documento se considera relevante para una consulta determinada si su contenido satisface la información representada por la consulta. Para determinar la relevancia, un evaluador humano revisa un documento y le asigna un valor de relevancia. El valor de relevancia puede ser binario ("relevante" o "no pertinente") o clasificadas (por ejemplo, "perfecto", "excelente", "bueno", "regular", aceptable", "no relevante", "nocivo").

4.7.1. Método de Evaluación

Un sistema de información se encuentra conformado por un conjunto de documentos y un determinado proceso de recuperación. Para la evaluación de la eficacia del sistema en una determinada consulta se necesita comparar los documentos que el sistema extrae con los documentos que el sistema cuenta sobre el tema consultado y son relevantes. La eficacia del sistema se desprenderá de la eficacia de los resultados de las distintas preguntas de los usuarios. Debemos indicar, asimismo, que no analizamos la evaluación de sistemas interactivos, por el contrario, solamente se estudian los aspectos relacionados con sistemas que podemos denominar por lotes, es decir, sistemas en los que se plantea una consulta y se analizan los resultados de la misma, sin considerar las operaciones posteriores del usuario. Los elementos que se requieren para la evaluación de un sistema de RI, son:

- Un conjunto de documentos indexados por el sistema del motor de búsqueda.
- Un conjunto predefinido de preguntas que representarán las necesidades de información de los usuarios, y se representarán en el sistema mediante expresiones de búsqueda con la sintaxis que se considere adecuada, según las características del buscador.
- Un conjunto de documentos relevantes, necesarios para las medidas y los parámetros de eficacia, que respondan a las preguntas correspondientes.

Entonces, los experimentos a realizar consisten en evaluar el rendimiento del buscador Google en un sitio web construido al efecto con documentos de un corpus

estándar. Para buscar las páginas web (documentos relevantes) respecto a una consulta, utilizamos la versión avanzada de Google en línea. Le preguntamos al buscador Google las consultas (los temas del corpus) y guardamos las páginas web recuperadas. Después se compara los resultados obtenidos con los documentos relevantes del corpus utilizando las medidas de evaluación como la precisión y la exhaustividad. Estas medidas serán comparadas con los valores de referencia para luego sacar conclusiones.

4.7.2. Colección de Prueba

Como hemos indicado en detalle (ver Sección 3.2.4), la colección de documentos Hamshahri es el único corpus de referencia de documentos persa que es estándar y está construido según las especificaciones de *TREC* (AleAhmad, et al., 2009). En esta parte de nuestros experimentos hemos elegido la versión 1 de *CLEF2008* de la misma colección. Porque esta versión tiene más número de temas (100 *topics*) y además, los documentos están en formato XML y la conversión de ellos a un formato HTML será más fácil. En la siguiente sección se explica el proceso de conversión de documentos. Las características del corpus Hamshahri (versión 1 de *CLEF2008*) que se utiliza en estos experimentos están indicadas en la Tabla 4.4.

Tabla 4.4: Las características del corpus Hamshahri

Atributos	Valores
Tamaño de colección	700 MB
Formato de documentos	XML
Número de documentos	166.774
Número de términos únicos	417.339
Longitud media de documentos	380 términos
Número de categorías	82
Número de temas	100
Formato de temas	XML

4.7.3. Construcción del Sitio Web con Documentos Persas

Esta primera etapa consiste en construir un sitio web con documentos del corpus Hamshahri pero en un formato de HTML. El corpus contiene 166.774 documentos en formato XML en un total de 1.927 archivos. Cada archivo tiene en promedio 87 documentos y consta con etiqueta abierta y cerrada de <HAMSHAHRI>. Cada documento se comienza con etiqueta <DOC> et el propio contenido textual del documento está rodeado por la etiqueta <TEXT>. Hay otras etiquetas como por ejemplo, la fecha, la categoría y etc. Como cada archivo XML del corpus tiene más de 85 documentos tenemos, en primer lugar, que dividir los archivos de XML para obtener los documentos separados de tal manera que cada documento sea en un archivo de formato XML. Para esta tarea utilizamos un programa en Java que permite extraer cada documento del archivo inicial y ponerlo en un archivo separado. Este programa va, entonces, a separar los documentos de cada archivo y recomienza de nuevo con el siguiente archivo hasta que no quede más archivo en la carpeta. Añadimos a cada archivo XML obtenido la etiqueta abierta y cerrada de <HAMSHAHRI>. Al final de esta operación tenemos 166.774 archivos XML separados.

Tomamos nota de que en el archivo de juicio de los documentos relevantes que está disponible con el corpus, cada documento (relevante o no relevante) se muestra por su número de <DOCID> o <DOCNO>. La elección de este nombre para el archivo obtenido nos ayudará a ver los documentos que serán seleccionados por el motor de búsqueda. Esto nos permitirá hacer una comparación entre los resultados obtenidos por el motor de búsqueda y los documentos relevantes ya mencionados en el archivo de juicio. Entonces, utilizamos otro programa en Java que nos permite elegir para cada archivo el nombre de DOCID o DOCNO (<DOCID> tiene el mismo atributo que <DOCNO>) del mismo documento que lo podemos leer en la etiqueta <DOCID> o <DOCNO>.

En la Figura 4.2 podemos ver un archivo XML de Hamshahri después de haber sido extraído del archivo inicial. El nombre del archivo es “H-750402-6S1.xml” como el valor de su etiqueta <DOCID>. Los detalles de documentos obtenidos, después

de haber separados de los archivos iníciales, están indicados según el año de publicación en la Tabla 4.5.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE HAMSHAHRI SYSTEM "hamshahri.dtd">
<HAMSHAHRI>
<DOC>
<DOCID>H-750402-6S1</DOCID>
<DOCNO>H-750402-6S1</DOCNO>
<DATE>1996-06-22</DATE>
<CAT xml:lang="fa">علمی فرهنگی</CAT>
<CAT xml:lang="en">Science and Culture</CAT>
<TEXT>
احتمال اعتصاب آموزگاران در جمهوري
آذربايجان
به گفته يك نماينده مجلس ملي جمهوري آذربايجان: اگر حقوق آموزگاران
افزايش نيابد احتمالا آنان اعتصاب خواهند كرد.
بابا خان مراداف، كه در اجلاس مجلس ملي سخن مي گفت; وضعيت زندگي
آموزگاران اين جمهوري را بسيار دشوار خواند.
به گزارش خبرگزاري توران ازجمهوري آذربايجان، مراداف تاكيد كرد
كه تنها در شهر بيش باكو از دوهزارو ۳۰۰ نفر ازآموزگاران دست از
كاركشيده اند.
رسول قلي اف، رئيس مجلس ملي جمهوري آذربايجان، حل اين مشكل توسط مجلس
ملي اين جمهوري را غيرممكن خواند.
وي تاكيد كرد كه در بودجه دولت، افزايش حقوقي براي معلمان پيش بيني
نشده است. درحال حاضر، حقوق متوسط آموزگاران در اين جمهوري، حدود
هزار ۳۵ منات /۷ ۵معادل دلار است.
</TEXT>
</DOC>
</HAMSHAHRI>
```

Figura 4.2: Un documento del corpus Hamshahri en formato XML

La segunda operación consiste en convertir todos estos archivos en archivos de formato HTML. El contenido de la parte cuerpo del archivo HTML es el mismo que el valor de la etiqueta <TEXT> del archivo XML. Para esta operación utilizamos un programa de Java con la ayuda del lenguaje XSL que nos permite diseñar la estructura del archivo HTML.

Tabla 4.5: Nº de documentos de Hamshahri en XML según el año de publicación

Año de publicación	Nº archivos iníciales en XML	Nº archivos finales en XML
1996	156	12.516
1997	286	23.883
1998	282	21.935
1999	278	23.387
2000	287	24.061
2001	269	22.211
2002	334	35.135
2003	35	3.646
Total	1.927	166.774

4.7.3.1. Lenguaje XSL

XSL son las siglas de "*Extensible Stylesheet Language*" (lenguaje extensible de hojas de estilo), una familia de lenguajes basados en el estándar XML que permite describir cómo debe ser transformada o formateada la información contenida en un documento XML para su presentación en un medio (generalmente, una página Web). XSL está formado por tres especificaciones recomendadas oficialmente por el W3C[22]:

- XSLT, *Extensible Stylesheet Language Transformations* (lenguaje de hojas extensibles de transformación) que permite convertir documentos XML de una sintaxis a otra (por ejemplo, de un XML a otro o a un documento HTML).
- XSL-FO, *Extensible Stylesheet Language Formatting Objects* (lenguaje de hojas extensibles de formateo de objetos", que permite especificar el formato visual con el cual se quiere presentar un documento XML (usado principalmente para generar documentos PDF).

[22] El World Wide Web Consortium, abreviado W3C, es un consorcio internacional que produce recomendaciones para la World Wide Web.

- XPATH o XML *Path Language*, que es una sintaxis (no basada en XML) para acceder o referirse a porciones de un documento XML.

Hamshahri corpus document

DOC ID : H-750402-6S1

Date of Document: 1996-06-22

احتمال اعتصاب آموزگاران در جمهوري آذربايجان به گفته يك نماينده مجلس ملي جمهوري آذربايجان: اگر حقوق بابا خان مراداف، كه در اجلاس مجلس ملي سخن مي .آموزگاران افزايش نيابد احتمالا آنان اعتصاب خواهند كرد به گزارش خبرگزاري توران ازجمهوري .گفت; وضعيت زندگي آموزگاران اين جمهوري را بسيار دشوار خواند آذربايجان، مراداف تاكيد كرد كه تنها در شهر بيش باكو از دوهزارو ۳۰۰ نفر ازآموزگاران دست از كاركشيده رسول قلي اف، رئيس مجلس ملي جمهوري آذربايجان، حل اين مشكل توسط مجلس ملي اين جمهوري را .اند وي تاكيد كرد كه در بودجه دولت، افزايش حقوقي براي معلمان پيش بيني نشده است. درحال .غيرممكن خواند حاضر، حقوق متوسط آموزگاران در اين جمهوري، حدود هزار ۳۵ منات /۷ ۵معادل دلار است.

Figura 4.3: El mismo documento de la Figura 4.2 en formato HTML

Después de esta operación tenemos 166.774 archivos HTML. El nombre de cada archivo queda el mismo nombre que el archivo XML pero con extensión “html”. En la parte B del Apéndice se puede ver los programas de Java y el archivo XSL que nos permite diseñar nuestras páginas web. En la Figura 4.3 tenemos el mismo documento citado más arriba en formato HTML llamado “H-750402-6S1.html”.

4.7.3.2. Registro del Sitio Web

Con ayuda de la infraestructura del departamento de informática de la universidad de Valladolid creamos nuestro sitio web llamado “http://farsidoc.infor.uva.es”. En el archivo “*robots.txt*” indicamos que todos los robots pueden rastrear todas las carpetas de documentos disponibles. Un archivo *robots.txt* es un archivo que se encuentra en la raíz de un sitio e indica a qué partes no quieres que accedan los rastreadores de los motores de búsqueda. El archivo utiliza el “estándar de exclusión de robots”, que es un protocolo con un pequeño conjunto de comandos que se puede utilizar para indicar el acceso al sitio por sección y por tipos específicos de rastreadores web.

Para construir el archivo *sitemap*, seguimos las indicaciones de Google según las cuales cada archivo *sitemap* debe tener al máximo 50.000 URL y no debe ser más de 10 MB. Un *sitemap* es un archivo en el que se pueden enumerar las páginas web del sitio propio para informar a Google y a otros motores de búsqueda sobre la organización del contenido del sitio. Los rastreadores web de los motores de búsqueda, por ejemplo, *Googlebot*, leen este archivo para rastrear el sitio de forma más inteligente. En nuestro caso hemos tenido que hacer múltiplos *sitemaps*. Cada *sitemap* es para los documentos de cada año de publicación (del 1996 hasta 2003). El *sitemap* principal que es el archivo indexado de todos los *sitemaps*. Los detalles de los archivos *sitemap.xml* y *robots.txt* están disponibles en la misma localidad.

4.7.4. Indexación de Páginas por Google

Una vez preparadas las páginas web pasamos a la siguiente etapa que es la indexación de las páginas web por Google. Como el nombre de las enlaces de indexación es demasiado mucho (166.774 páginas web), entonces la carga de la página inicial es muy lento. Por eso, hemos decidido separar los documentos según el año de sus publicaciones. Así, podemos repartir los documentos en diferentes carpetas y tener una página web de indexación más pequeña para cada carpeta. Hacemos 8 carpetas cada uno corresponde a un año del 1996 hasta 2003. Al final, todos los documentos fueron subidos en nuestro sitio web. Una vez cargando todos los elementos solicitamos que Google indexara el contenido del sitio web. Después de haber esperado acerca de 10 días hasta que Google procese la solicitud y rastree e indexe las páginas resulta que Google sólo había indexado 41.340 páginas, es decir, menos de 25% de todos los documentos. Tras varios intentos, el resultado de indexación era el mismo.

Según las indicaciones de *Google Webmaster Tools*, Google no garantiza indexar todas las páginas web de un sitio web en particular. En nuestro caso las razones pueden ser o bien el tamaño del sitio web es muy grande o el contenido del sitio web no es relevante o útil para los usuarios. Además, no tenemos ninguna indicación de que

podemos saber qué páginas están indexadas por Google para poder continuar nuestros experimentos considerando sólo las páginas indexadas.

Para resolver este problema, hemos distribuido los documentos en cuatro diferentes sitios web denominados “farsidoc1”, “farsidoc2”, “farsidoc3” y “farsidoc4” para disminuir el tamaño de documentos. Cada nuevo sitio web representa los documentos de dos años de publicación. Después de construir los nuevos archivos *sitemap* y *robot.txt* para cada sitio web y hacer la indexación de las páginas, hemos subido los documentos en sus sitios web. Solicitamos de nuevo a Google que analice nuestros sitios web. En la Tabla 4.6 se resume la repartición de documentos en cada sitio web y el número de páginas indexadas por Google.

Tabla 4.6: Repartición de documentos en diferentes sitios web

Sitio web	Nº total de documentos	Año de publicación	Nº de doc. indexados por Google	% doc. indexados
http://farsidoc.infor.uva.es	166.774	1996-2003	41.340	24.8%
http://farsidoc1.infor.uva.es	36.399	1996-1997	33.908	93.2%
http://farsidoc2.infor.uva.es	45.322	1998-1999	5.646	12.5 %
http://farsidoc3.infor.uva.es	46.272	2000-2001	10.107	21.9 %
http://farsidoc4.infor.uva.es	38.781	2002-2003	22.781	58.7 %

Teniendo en cuenta el número de páginas indexadas en cada sitio web, así que decidimos hacer una evaluación del primer sitio web, es decir, “farsidoc1”. Como podemos ver más de 93% de documentos de este sitio web están indexados por Google y por lo tanto nuestra evaluación puede ser más significativa con un menor margen de error que en el resto de casos.

En este caso, hemos extraído del archivo de juicio de relevancia todas las informaciones correspondientes a los documentos que están localizados en el primer sitio Web “farsidoc1”. Este nuevo archivo de juicio de relevancia deber tener, según las especificaciones de TREC, la forma siguiente:

Query-number	0	document-id	relevance

donde "*query-number*" es el numero de la consulta, "*document-id*" es un externa identificador para el documento recuperado que es el nombre del documento, "0" es un valor constante y el valor de "*relevance*" es la relevancia asignado a un documento por una consulta particular. La relevancia es 0 (no relevante) o bien 1 (relevante). Recordamos que el valor de "*query-number*" en el corpus Hamshahri es de 551 hasta 600 (las primeras 50 temas creados para Hamshahri en la *CLEF* 2008) y de 601 hasta 650 (las segundas 50 temas creados para Hamshahri en la *CLEF* 2009). El valor de "*document-id*" es el identificador del documento que, en nuestro caso, es el valor de la etiqueta <DOCID> o bien <DOCNO>. Los archivos de juicio de relevancia obtenidos están disponibles en el sitio web "farsidoc1".

- qrel_ham2008.test: es el archivo de juicio de relevancia para las 50 primeras consultas de CLEF2008.
- qrel_ham2009.test: es el archivo de juicio de relevancia para las 50 primeras consultas de CLEF2009.

4.7.5. Resultados de Búsqueda por Google

Las búsquedas se realizaron durante los meses de febrero y marzo 2014. El proceso de búsqueda fue realizado mediante la versión de búsqueda avanzada de Google. Entonces, el texto de las consultas introducido en el cuadro de búsqueda de Google estuvo exactamente el mismo texto del título de los temas (*topics)* del corpus Hamshahri y el domino o sitio era "http://farsidoc1.infor.uva.es". Recordamos que los archivos de consultas son los siguientes y están todos disponibles en nuestro sitio Web.

- Persian_topics_CLEF2008.xml: es el archivo que contiene las 50 consultas en la lengua persa de CLEF2008.
- English_topics_CLEF2008.xml: es el archivo que contiene las 50 consultas traducidas en inglés de CLEF2008.
- Persian_topics_CLEF2009.xml: es el archivo que contiene las 50 consultas en la lengua persa de CLEF2009.

- English_topics_CLEF2009.xml: es el archivo que contiene las 50 consultas traducidas en inglés de CLEF2009.

Una vez obtenido los resultados de búsqueda para cada consulta que es la lista ordenada de *URL* de los documentos recuperados, necesitaremos guardar estos resultados en unos archivos para las próximas operaciones de comparaciones. El registro de los resultados de búsqueda se hace mediante "SEOQuake" que es una herramienta específica para esta tarea.

4.7.5.1. Herramienta SEOQuake

SeoQuake[23] es una extensión de Mozilla Firefox de *SEO* (*Search Engine Optimization*) dirigida principalmente a ayudar a los administradores de sitios Web que se ocupan de la optimización de los motores de búsqueda y la promoción de Internet de sitios Web. SeoQuake permite obtener e investigar muchos parámetros de *SEO* importante del proyecto en estudio sobre la marcha, guardarlos para el trabajo futuro, compararlos con los resultados obtenidos para otros proyectos competitivos. Después de la presentación de la consulta en el motor de búsqueda el usuario se presenta *SERP* (*Search Engine Result Pages*) con los resultados de búsqueda. *SeoQuake* muestra los valores de los parámetros para los resultados de la búsqueda en cada resultado de la búsqueda aparecerá en *SERP*. El conjunto de parámetros que se muestra es completamente personalizable por el usuario. La carga de parámetros se puede realizar de dos maneras diferentes: al mismo tiempo con la carga de la *SERP* o después de cargar *SERP*, bajo demanda de los usuarios. Las funciones disponibles son el *ranking* de los resultados en orden ascendente/descendente con el parámetro seleccionado y el almacenamiento de resultados en el archivo.

El resultado obtenido por el buscador se puede guardar en un archivo de formato *CSV*. Los archivos *CSV* (del inglés *Comma Separated Values*) son un tipo de

[23] http://www.seoquake.com/

documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas y las filas por saltos de línea. El formato es utilizado en muchos programas de bases de datos, hojas de cálculo y gestores de contactos para almacenar listas de información. Como es un archivo de texto, el formato es ampliamente compatible.

Para cada consulta, tenemos como resultado de la búsqueda la *URL* del documento buscado, el valor de ranking y de la similitud. El valor de ranking es el número de la aparición del documento según el orden visto por Google. El valor de similitud es igual a 0 puesto que la búsqueda se hace en un sitio web determinado. Guardamos este resultado en un archivo de formato *CSV*. Para que este obtenido archivo sea compatible con las normas requeridas de TREC, hacemos unas modificaciones de siguientes maneras:

Para cada consulta, agregamos a cada fila del archivo de resultados el número de la consulta usando por el archivo de juicio de relevancia. Otra modificación es convertir la *URL* obtenido por el título del documento que corresponde al "*document-id*" que en nuestro caso es el valor de <DOCID> o <DOCNO>. Finalmente, nuestro archivo de resultado es un archivo de texto y tiene la forma siguiente:

Query-number Q0 document-id rank score run-id

donde "*query-number*" es el numero de la consulta, "*document-id*" es el identificador del documento recuperado y "*score*" es el valor de similitud. "Q0" (Q cero) y "*run-id*" son dos constantes que se usan por algunos programas de evaluación. Debe haber un espacio entre cada elemento de una línea de resultados. Finalmente, tenemos dos archivos estructurados de resultados de búsqueda por Google que están disponibles en nuestro sitio Web "farsidoc1" y son:

- results_ham2008.test: es el archivo de resultado de búsqueda por Google estructurado según TREC para las consultas de CLEF2008.
- results_ham2009.test: es el archivo de resultado de búsqueda por Google estructurado según TREC para las consultas de CLEF2009.

4.8. Medidas de Evaluación

Las medidas de evaluación aquí descritas pueden calcularse bien como medidas puntuales relativas a una consulta concreta, bien como medidas globales relativas a un conjunto de consultas. En este último caso las medidas son calculadas promediando los valores obtenidos para cada consulta individual respecto al número de consultas empleado, salvo en el caso de la precisión media de documento, que será descrito en detalle más adelante.

4.8.1. Precisión y Exhaustividad

El método más habitual para medir la calidad de un sistema es la utilización de diagramas precisión-exhaustividad (*precisión-recall*) (Baeza-Yates, et al., 1999). Tomemos una consulta concreta de la colección de pruebas y el conjunto de documentos relevantes para dicha consulta. Sea |*b*| número de documentos relevantes. Apliquemos esa consulta al sistema que se desea evaluar. Para esa consulta se ha recuperado un conjunto de documentos. Sea |*a*| el número de documentos recuperados y denominamos |*c*| el número de documentos recuperados que son relevantes. La Figura 4.4 ilustra estos conjuntos. Las medidas de precisión y exhaustividad se definen como:

Precisión: se define como la proporción de los documentos recuperados que son relevantes y permite evaluar la habilidad del sistema para ordenar primero la mayoría de los documentos relevantes.

$$Precisión = \frac{|c|}{|a|}$$

Exhaustividad: se define como la proporción de los documentos relevantes que han sido recuperados y permite evaluar la habilidad del sistema para encontrar todos los documentos relevantes de la colección.

$$Exhaustividad = \frac{|c|}{|b|}$$

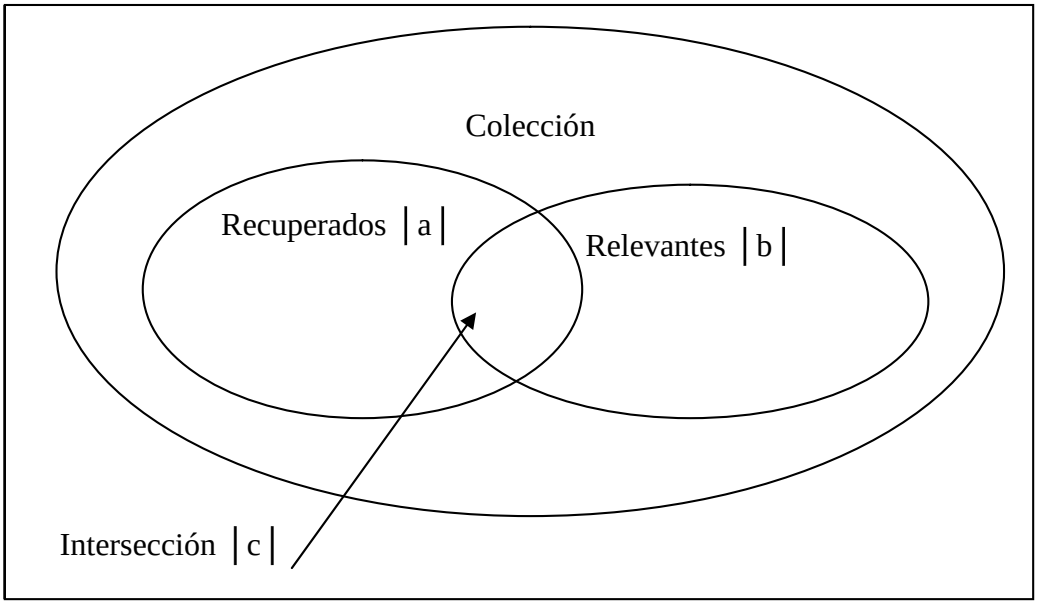

Figura 4.4: Precisión-exhaustividad para una consulta de ejemplo

Estas medidas de precisión y exhaustividad suponen que se han examinado todos los documentos recuperados. Sin embargo, al usuario normalmente no se le presentan todos los documentos a la vez, sino que el sistema se los presenta ordenados utilizando algún criterio interno. El usuario entonces empieza a examinar la lista de documentos ordenados, empezando por el primero. En este proceso las medidas de precisión y exhaustividad van variando con cada uno de los documentos examinados. Para la explicación se ha seguido más o menos fielmente (Baeza-Yates, et al., 1999).

En nuestros experimentos empleamos medidas que miden tanto la precisión como la relación entre precisión y exhaustividad. Las medidas utilizadas son las siguientes: Diagrama precisión-recuperación interpolada, precisión a los n documentos devueltos (donde n= 5 y 10 y…), R-precisión y precisión media de documento (*Mean Average Precision*, *MAP*). Los valores de todas estas medidas van a ser obtenidos mediante la herramienta llamada TREC_Eval[24], de uso ampliamente difundido en este tipo de investigaciones.

[24] http://trec.nist.gov/trec_eval

4.8.2. Diagrama Precisión-Recuperación Interpolada

Resulta mucho más interesante obtener una medida que involucre varias consultas. Normalmente se toma interpolación de la precisión para 11 puntos estándar de recuperación en los niveles del 0%, 10%, 20%,...,100% (Manning, et al., 2008). La precisión para cada uno de estos niveles se calcula como el máximo valor de precisión entre ese valor y el siguiente. Formalmente, podemos definirlo de esta manera, sea n_j, con $j \in \{0, 1, 2,..., 10\}$, el nivel estándar j-ésimo de recuperación, entonces, el valor de precisión interpolada para ese nivel viene dado como:

$$P(n_j) = \max_{n_j \le n \le n_{j+1}} P(n)$$

Puesto que el valor de precisión para cada nivel de recuperación interpolada se calcula sobre el valor máximo de precisión entre un nivel y el siguiente (por ejemplo, para calcular el valor interpolado de precisión para el nivel del 50% se observan los valores de precisión en el rango entre el 50% y el 60%), siempre se debe calcular primero el valor para el 100%.

4.8.3. Precisión Media no Interpolada

Una de las medidas más comunes en la comunidad del *TREC* es *Mean Average Precision* (*MAP*), la cual genera un único valor que resume el rendimiento de un sistema a distintos niveles de exhaustividad (Manning, et al., 2008). Además, esta medida ha mostrado tener un buen poder de discriminación[25] y una buena estabilidad[26].

Cuando se realiza la evaluación utilizando *MAP*, para cada consulta se calcula la media de los valores de precisión obtenidos cada vez que se encuentra un documento

[25] Cuanto más discriminativa es una medida, menos empates habrá entre sistemas y menor será la diferencia necesaria para concluir qué sistema es mejor (Buckley, et al., 2000)

[26] La estabilidad es el error asociado a la conclusión el sistema A es mejor que el sistema B (Buckley, et al., 2000)

relevante. El valor final para el conjunto de consultas es la media de los valores calculados para cada consulta. Es decir, si el conjunto de documentos relevantes para una consulta $q_j \in Q$ es $\{d_1,...,d_{mj}\}$ y R_{jk} es el conjunto de documentos recuperados ordenados desde el primero hasta el documento d_k, entonces se tiene que la fórmula de *MAP* es la definida en la ecuación 4.1.

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \qquad (4.1)$$

Para una única consulta, *MAP* aproxima el área bajo la curva *precisión-exhaustividad*, mientras que dado un conjunto de consultas *MAP* aproxima el área bajo la curva *precisión- exhaustividad* de dicho conjunto de consultas. Además, mediante el uso de la media aritmética en *MAP* cada consulta tiene el mismo peso en el valor final (Rodrigo Yuste, 2010).

4.8.4. Precisión Media a Ciertos Documentos Relevantes Vistos

Se trata de obtener un valor medio según se van considerando los documentos relevantes, para el total de documentos vistos. Es decir la precisión (porcentaje de documentos recuperados que son relevantes) después de n documentos recuperados que sean relevantes o no relevantes (Manning, et al., 2008). Los valores están promediados sobre todas las consultas. Si no se recuperaron n documentos para una consulta entonces todos los documentos que faltan se suponen que no son relevantes. Esta medida favorece aquellos sistemas que obtienen antes los documentos relevantes. Pero no debe despistarnos, pues un valor alto puede enmascarar valores muy bajos de recuperación (por ejemplo, si solamente se recuperan dos documentos relevantes en las posiciones primera y segunda, el valor de esta medida sería alto, pero la recuperación sería pequeña).

4.8.5. R-Precisión

Es la precisión obtenida a los R documentos devueltos, donde R es el número de documentos relevantes para esa consulta (Manning, et al., 2008). Esta medida individual permite observar el comportamiento del algoritmo de recuperación para todas las consultas de la colección de prueba. También se puede obtener una R−Precisión media para toda la colección de consultas, pero puede ser una media poco caracterizadora del sistema.

Tabla 4.7: La lista de archivos publicados en sitio web "farsidoc1"

Archivo	Descripción
qrel_ham2008.test	Juicio de relevancia para 50 consultas de CLEF2008
Persian_topics_CLEF2008.xml	Texto de 50 consultas de CLEF2008 en persa
English_topics_CLEF2008.xml	Texto de 50 consultas de CLEF2008 traducidas en inglés
results_ham2008.test	Archivo de resultados de búsqueda por Google estructurado según TREC para las consultas de CLEF2008
results_all_queries2008	Medidas de precisión para cada consulta y el conjunto de consultas de CLEF2008
qrel_ham2009.test	Juicio de relevancia para 50 consultas de CLEF2009
Persian_topics_CLEF2009.xml	Texto de 50 consultas de CLEF2009 en persa
English_topics_CLEF2009.xml	Texto de 50 consultas de CLEF2009 traducidas en inglés
results_ham2009.test	Archivo de resultados de búsqueda por Google estructurado según TREC para las consultas de CLEF2009
results_all_queries2009	Medidas de precisión para cada consulta y el conjunto de consultas de CLEF2009.

4.9. Resultados Obtenidos

Las experimentas de la evolución se hacen en dos etapas. En la primera etapa se considera las 50 consultas creadas para el corpus Hamshahri en la conferencia de *CLEF* 2008 y la segunda concierne las 50 consultas creadas para el corpus en la *CLEF* 2009, en total 100 consultas. Los resultados obtenidos son en forma de tablas de resumen en las que aparecen, primero resultados individuales para cada consulta y al final la media para todas las consultas del experimento. Es habitual encontrarse datos como el número de documentos relevantes, número de documentos relevantes recuperados, valores

interpolados de precisión, R−Precisión, precisión a cierta cantidad de documentos vistos, etc. Los detalles de resultados obtenidos están publicados en nuestro sitio web "farsidoc1" y para facilitar al lector recapitulamos y agrupamos todos los archivos utilizados y obtenidos en la Tabla 4.7.

El resumen de los valores obtenidos para el conjunto de 50 consultas de CLEF2008 se muestra en la Tabla 4.8. La primera columna es el identificador de la consulta, la segunda columna es el número de documentos recuperados por Google, La tercera columna nos indica el número de todos los documentos relevantes que están en el corpus Hamshahri según datos del archivo de juicio de relevancia. La cuarta columna es el número de documentos que están recuperados por Google y que son también relevantes. Otras columnas están dedicadas a los valores de precisión media y R-Precisión exacta.

Tabla 4.8: Resumen de medidas de precisión para 50 consultas de CLEF2008

N° de consulta	N° de doc. recuperados	N° de doc. relevantes	N° de documentos recuperados-relevantes	Precisión media	R-precisión exacta
551	11	22	10	0,4504	0,4545
552	38	22	9	0,1048	0,2273
553	-	-	-	-	-
554	18	20	1	0,0033	0,0500
555	6	18	1	0,0556	0,0556
556	4	16	0	0,0000	0,0000
557	-	-	-	-	-
558	1	2	0	0,0000	0,0000
559	19	8	6	0,2989	0,1250
560	115	43	24	0,0970	0,1163
561	3	52	2	0,0385	0,0385
562	1	1	0	0,0000	0,0000
563	6	4	1	0,2500	0,2500
564	20	13	9	0,3941	0,3846
565	24	3	2	0,1503	0,3333
566	11	6	1	0,0833	0,1667
567	107	34	19	0,1032	0,1176

Tabla 4.8 (continuación)

N° de consulta	N° de doc. recuperados	N° de doc. relevantes	N° de documentos recuperados-relevantes	Precisión media	R-precisión exacta
568	33	23	13	0,2897	0,4348
569	9	6	0	0,0000	0,0000
570	55	41	9	0,0846	0,1951
571	126	12	8	0,0429	0,0833
572	6	6	4	0,4611	0,6667
573	-	-	-	-	-
574	-	-	-	-	-
575	-	-	-	-	-
576	142	29	19	0,1289	0,1379
577	14	47	5	0,0720	0,1064
578	21	41	8	0,0846	0,1951
579	36	13	7	0,0851	0,0000
580	115	101	41	0,1124	0,3663
581	119	79	48	0,2705	0,4810
582	63	61	21	0,1197	0,3279
583	73	63	29	0,2354	0,4286
584	51	53	16	0,0781	0,3019
585	121	19	12	0,0538	0,0526
586	22	9	6	0,2100	0,3333
587	5	4	1	0,0500	0,0000
588	35	16	8	0,1797	0,2500
589	75	29	20	0,1390	0,1379
590	36	42	12	0,1161	0,2857
591	105	51	31	0,2006	0,3333
592	5	9	2	0,1296	0,2222
593	3	1	1	0,3333	0,0000
594	3	4	2	0,2917	0,5000
595	74	6	2	0,0521	0,1667
596	12	18	5	0,1364	0,2778
597	80	31	21	0,2381	0,2903
598	-	-	-	-	-

Tabla 4.8 (continuación)

N° de consulta	N° de doc. recuperados	N° de doc. relevantes	N° de documentos recuperados-relevantes	Precisión media	R-precisión exacta
599	88	23	11	0,0575	0,0435
600	111	22	15	0,1982	0,2727
Total	2.022	1.123	462	0,1473	0,2093

Desde los datos de la Tabla 4.8 podemos ver que sobre los 1.123 documentos relevantes para el conjunto de las consultas, Google ha encontrado 462 documentos relevantes. Es decir, sólo 41% de los documentos relevantes son recuperados por Google. Una pequeña porción de los 59% documentos no recuperados podría justificarse por el hecho de que una pequeña cantidad de páginas web todavía no están indexadas por Google. Pero, esto no explica que casi más de 50% de los documentos relevantes no se hayan podido recuperar. Entre las 50 consultas, 6 consultas no están consideradas en el análisis sobre el conjunto de las consultas. Esto puede explicarse que Google no ha recuperado ninguna página web (que sea relevante o no para una dada consulta) o puede que no haya documentos relevantes en el conjunto de documentos entre los años 1996 y 1997, según el archivo de juicio de relevancia.

De la misma manera, el resumen de los valores obtenidos para el conjunto de 50 consultas de CLEF2009 se muestran en la Tabla 4.9. Como nos indica estos datos sobre los 670 documentos relevantes para las 50 consultas, Google ha encontrado 83 documentos relevantes. Es decir, sólo 12% de los documentos relevantes son recuperados por Google. Al igual que sucedió antes con CLEF2008, una pequeña porción de los 88% documentos no recuperados puede justificarse por el hecho de que una pequeña cantidad de páginas todavía no es indexada por Google. Pero, esto no explica que casi más de 80% de los documentos relevantes no se han podido recuperar.

Tabla 4.9: Resumen de medidas de precisión para 50 consultas de CLEF2009

N° de consulta	N° de doc. recuperados	N° de doc. relevantes	N° de documentos recuperados-relevantes	Precisión media	R-precisión exacta
601	35	21	3	0,0089	0,0000
602	5	27	1	0,0370	0,0370
603	-	-	-	-	-
604	13	23	0	0,0000	0,0000
605	2	9	0	0,0000	0,0000
606	3	42	2	0,0397	0,0476
607	-	-	-	-	-
608	30	17	3	0,0338	0,1176
609	3	14	1	0,0714	0,0714
510	-	-	-	-	-
611	-	-	-	-	-
612	-	-	-	-	-
613	16	31	5	0,1427	0,1613
614	-	-	-	-	-
615	-	-	-	-	-
616	10	31	9	0,2835	0,2903
617	2	6	1	0,1667	0,1667
618	-	-	-	-	-
619	17	29	6	0,1099	0,2069
620	-	-	-	-	-
621	2	18	0	0,0000	0,0000
622	12	8	0	0,0000	0,0000
623	3	19	1	0,0175	0,0526
624	2	8	1	0,0625	0,1250
625	1	6	1	0,1667	0,1667
626	18	20	3	0,0667	0,1500
627	19	21	11	0,3812	0,5238
628	-	-	-	-	-
629	1	18	0	0,0000	0,0000
630	1	20	0	0,0000	0,0000
631	26	16	2	0,0122	0,0625

Tabla 4.9 (continuación)

N° de consulta	N° de doc. recuperados	N° de doc. relevantes	N° de documentos recuperados-relevantes	Precisión media	R-precisión exacta
632	1	6	1	0,1667	0,1667
633	10	10	4	0,1383	0,4000
634	7	14	4	0,2298	0,2857
635	4	35	2	0,0571	0,0571
636	126	44	13	0,0275	0,0682
637	4	28	2	0,0595	0,0714
638	13	17	0	0,0000	0,0000
639	5	21	3	0,1429	0,1429
640	-	-	-	-	-
641	-	-	-	-	-
642	14	26	2	0,0099	0,0769
643	7	27	1	0,0074	0,0370
644	1	5	1	0,2000	0,2000
645	-	-	-	-	-
646	13	12	0	0,0000	0,0000
647	-	-	-	-	-
648	-	-	-	-	-
649	1	21	0	0,0000	0,0000
650	-	-	-	-	-
Total	427	670	83	0,0776	0,1084

4.10. Discusión de los Resultados Obtenidos

En esta sección analizamos los resultados obtenidos para sacar nuestras conclusiones. La Tabla 4.10 es la tabla de resumen de los resultados de medidas de precisión considerando el conjunto de consultas de CLEF2008 y la Tabla 4.11 es la de CLEF2009. Las primeras medidas son la precisión y exhaustividad interpoladas sobre el conjunto de consultas. Como hemos descrito antes, estas son dos medidas ampliamente aceptadas por la comunidad de RI y fueron planteadas por Cleverdon (Cleverdon, et al., 1966).

Tabla 4.10: Medidas de precisión de las consultas de CLEF2008

```
Queryid (Num):      44
Total number of documents over all queries
    Retrieved:    2022
    Relevant:     1123
    Rel_ret:       462
Interpolated Recall - Precision Averages:
    at 0.00       0.5313
    at 0.10       0.3797
    at 0.20       0.3063
    at 0.30       0.2278
    at 0.40       0.1949
    at 0.50       0.1377
    at 0.60       0.0909
    at 0.70       0.0161
    at 0.80       0.0076
    at 0.90       0.0076
    at 1.00       0.0076
Average precision (non-interpolated) for all rel docs (averaged over queries)
                0.1473
Precision:
  At    5 docs:   0.2409
  At   10 docs:   0.2114
  At   15 docs:   0.1924
  At   20 docs:   0.1841
  At   30 docs:   0.1606
  At  100 docs:   0.0973
  At  200 docs:   0.0525
  At  500 docs:   0.0210
  At 1000 docs:   0.0105
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.2093
```

Para comparar los resultados obtenidos con Google, elegimos los resultados obtenidos por unos experimentos realizados por Di Nunzio de la universidad de Padua de Italia (Di Nunzio, et al., 2008) como una base de referencia. Estos resultados son las conclusiones de una serie experimentos preparados y presentado oficialmente a la conferencia de CLEF (*Results for CLEF 2008 Adhoc Persian@CLEF Track*). Los experimentos se realizaron utilizando también la colección Hamshahri con el método de *polling*. El resultado de los experimentos realizados se encuentra en un archivo (AppendixB.pdf) que está disponible en nuestro sitio web “farsidoc1”.

Tabla 4.11: Medidas de precisión de las consultas de CLEF2009

```
Queryid (Num):       34
Total number of documents over all queries
    Retrieved:      427
    Relevant:       670
    Rel_ret:         83
Interpolated Recall - Precision Averages:
    at 0.00       0.5130
    at 0.10       0.3243
    at 0.20       0.1296
    at 0.30       0.0308
    at 0.40       0.0308
    at 0.50       0.0190
    at 0.60       0.0000
    at 0.70       0.0000
    at 0.80       0.0000
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.0776
Precision:
  At    5 docs:   0.2471
  At   10 docs:   0.1647
  At   15 docs:   0.1255
  At   20 docs:   0.0985
  At   30 docs:   0.0696
  At  100 docs:   0.0232
  At  200 docs:   0.0122
  At  500 docs:   0.0049
  At 1000 docs:   0.0024
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.1084
```

Para la primera comparación, hemos elegido uno de sus experimentos en el que no hay ninguna aplicación de lematizador sobre los documentos y el texto introducido para la búsqueda es el texto del título dentro de los campos de temas de las consultas del corpus (ver la página 176 del archivo AppendixB.pdf). Recordamos que hay tres campos para cada consulta (el titulo, la descripción y la narrativa) dentro del archivo de temas del corpus. Como se ha indicado antes, nosotros también hemos utilizado el texto del título como la consulta de búsqueda por Google. En la Tabla 4.12 tenemos los 11 valores de precisión interpolada para cada nivel de exhaustividad interpolada de los 3 experimentos.

Tabla 4.12: Valores de precisión y exhaustividad interpoladas obtenidos en 3 experimentos

Exhaustividad	Precisión Google CLEF2008	Precisión Google CLEF2009	Precisión Ad-hoc Persian@CLEF Track
0	0,5313	0,513	0,7839
0,1	0,3797	0,3243	0,656
0,2	0,3063	0,1296	0,5897
0,3	0,2278	0,0308	0,4671
0,4	0,1949	0,0308	0,3662
0,5	0,1377	0,019	0,1971
0,6	0,0909	0	0,0874
0,7	0,0161	0	0,0593
0,8	0,0076	0	0,0457
0,9	0,0076	0	0,021
1	0,0076	0	0,0048

En la Figura 4.5 mostramos la curva precisión-exhaustividad interpolada por todas las consultas para los dos experimentos con la de Di Nunzio (Di Nunzio, et al., 2008).

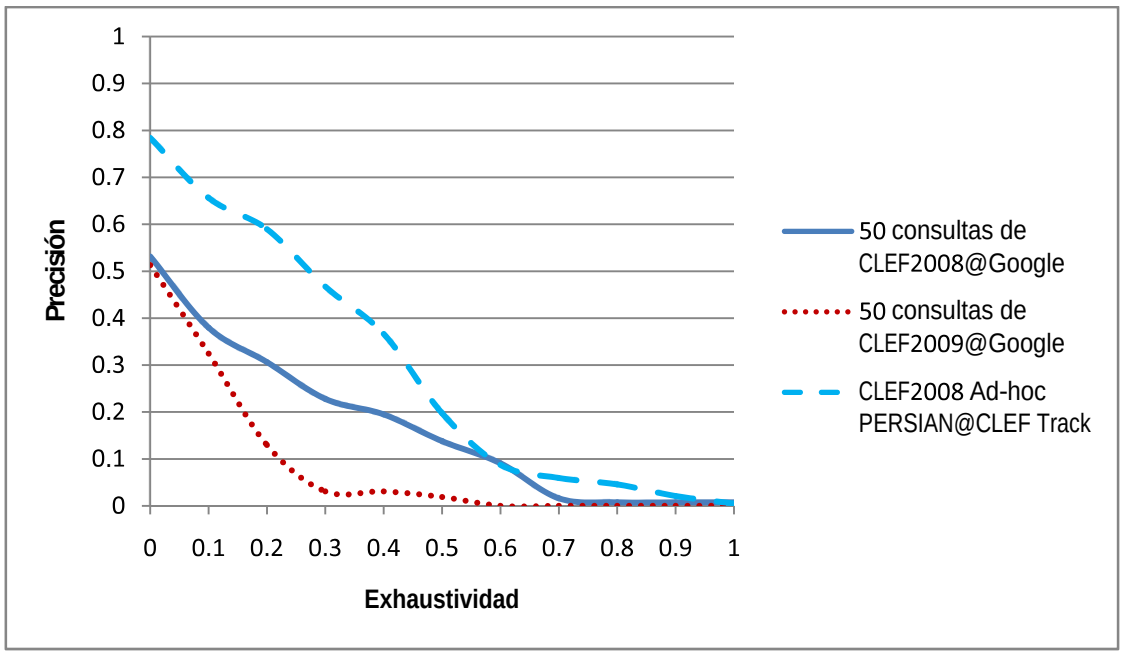

Figura 4.5: Diagrama precisión-exhaustividad interpolada para 3 experimentos

Como podemos observar en esta figura, a todos los niveles de exhaustividad, la precisión de la relevancia de los documentos recuperados por Google es menor que la de la base de referencia.

Otra comparación para evaluar el rendimiento de recuperación es el valor del promedio de media de precisión (*MAP*) basado en las 100 consultas. El uso de la media proporciona el mismo nivel de importancia para todas las consultas. En la Tabla 4.13 tenemos los valores de *MAP* para todos los experimentos.

Tabla 4.13: Medidas de *MAP* para los 3 experimentos

Valor	Experimentos con las consultas de TREC2008	Experimentos con las consultas de TREC2009	Experimentos de la base de referencia
MAP	0,1473	0,0776	0,2708

Comparando los valores obtenidos, tenemos un decremento de la precisión alrededor de 45,6% en el caso de los experimentos de TREC2008 y el 71,3% para los experimentos de TREC2009. Los resultados obtenidos describen la eficacia de Google y revelan que la precisión de los documentos devueltos por Google no es muy satisfactoria.

Si analizamos los resultados de precisión para cada consulta separadamente, podemos ver que las consultas número 551 y 572 tienen los mejores resultados de precisión. El texto de la consulta 551 es "تنیس جام ویمبلدون" (Copa de tenis de Wimbledon) y el de la consulta 572 es "کنسرت شجریان " (Concierto de Shayarián[27]). El texto de estas dos consultas contiene nombres propios. Existen otras consultas en las que hay dos nombres propios ایران (Irán) y تهران (Teherán). Aunque estas palabras son también nombres propios pero, debido a sus altas frecuencias en la colección, los

[27] Mohammad Reza Shayarián (23 de septiembre de 1940, Mashhad, Irán) es un cantante y compositor iraní de fama internacional. Ha sido calificado como «el mayor maestro vivo de la música tradicional persa Shayarián es conocido también como calígrafo y por su implicación en actividades caritativas.

documentos recuperados por las consultas que contienen estos términos no son muy relevantes. Recordamos que las dos palabras ocupan respectivamente 24º y 25º posiciones en la tabla de palabras claves del contenido construida por Google. Esto significa que el número de ocurrencia de estas palabras es muy alta en el conjunto de documentos del sitio web. Tenemos en la Tabla 4.14 los valores de *MAP* (el promedio medio de precisión) para las consultas 551 y 572.

Tabla 4.14: Valores de MAP para las consultas 551 y 572

	Promedio medio de precisión (*MAP*)	
Consultas	**Nuestros experimentos**	**Experimentos de la base de referencia**
Nº 551	0,4504	0,6526
Nº 572	0,4611	0,6924

Aquí, el decremento de la precisión es menos que sobre el conjunto de las consultas. En la consulta número 551 hay un decremento de 31% y 33,4% para la consulta número 572. En la Tabla 4.15 tenemos precisión media a ciertos documentos relevantes vistos para las dos consultas.

Tabla 4.15: Medidas de precisión por consultas 551 y 572 para los 100 primeros documentos recuperados

N° de documentos recuperados	Precisión consulta 551	Precisión consulta 572	Precisión de la base de referencia
5	1	0,6	0,592
10	0,9	0,4	0,576
15	0,6667	0,2667	0,568
20	0,5	0,2	0,554
30	0,3333	0,1333	0,53
100	0,1	0,04	0,3982

A partir de estos datos presentamos la curva de documentos recuperados vs la precisión media por los 3 experimentos (ver la Figura 4.6). Lo que se puede deducir de

las curvas es que en el caso de que se encuentren nombres propios en la consulta el resultado de búsqueda mejora en precisión, sobre todo, para los documentos recuperados entre 5 y 15 primeros. Generalmente, como los usuarios de internet se limitan a ver los primeros documentos recuperados por el motor de búsqueda entonces en este caso, la búsqueda de documentos relevantes será más precisa.

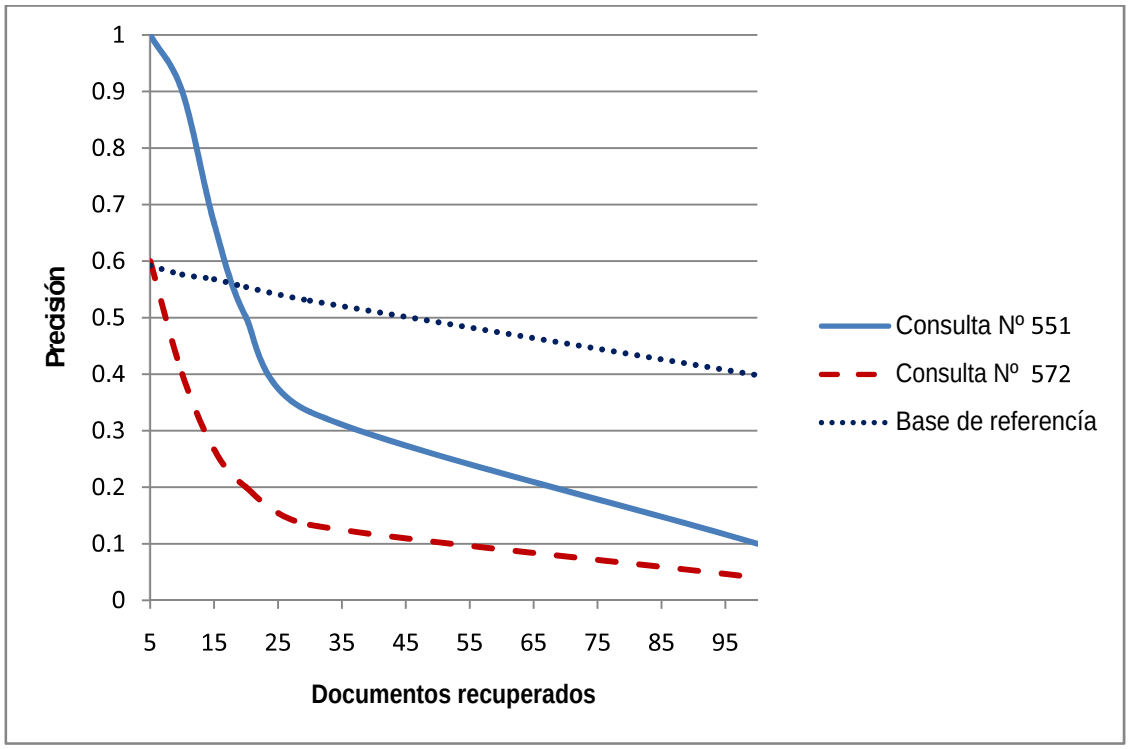

Figura 4.6: Diagrama de precisión vs documentos recuperados para las consultas 551 y 572

4.10.1. Índice de Google

Otro punto a considerar es la estrategia de Google para indexar las páginas web en persa. En la Tabla 4.16 ponemos las palabras claves del contenido indicadas por Google. La tabla está ordenada según la importancia (*significance*) de la palabra en los documentos del sitio web "farsidoc1". Según Google la importancia de una palabra es función de la ocurrencia de la palabra en la colección de los documentos. En esa tabla tenemos también las variantes encontradas por Google de cada palabra clave. Para

facilitar al lector, ponemos sólo las 20 primeras palabras con sus traducciones en español.

Tabla 4.16: Palabras claves del contenido en "farsidoc1"

Orden	Palabra clave	Traducción en español	Variantes encontradas
1	به	a	به, با, بي, بند, بدهد, بيم, بتون, بت, بنا
2	در	en	در, دره, درهم
3	از	desde	از
4	است	es	است, اي, اش, ام, ات, امان, اتو, اشان
5	مي	particula verbal del imperativo	مي, مه, متان, متون
6	كه	que	كه, كي, كاش, كيه
7	اين	esto, este	اين, اينان
8	را	râ (marcador del objeto)	را, راي, راست, راش, رايانه, رايي, رامون
9	هاي	sufio plural + *ezafe*	هاي, ها, هاست
10	آن	eso	آن, آنها, آنان, آنچه, آنجا, آنكه, آني, آنهم
11	كرد	hacer (derivado)	كرد, كرده, كردند, كردن, كردم, كرديم, كردي, كرديد, كردتو
12	شده	estar, ser (derivado)	شده, شدند, شدن, شدت, شد, شدگان, شدم, شديم, شدي
13	يك	uno	يك, يكي
14	كند	hacer (derivado)	كند, كنند, كنيد, كنيم, كنم, كني, كن, كنت, كنن, كنه, كنين
15	براى	para, por	براي, برا, برايان
16	بود	fue	بود, بودند, بوده, بودن, بودم, بودي, بوديم, بودو
17	شد	ser, estar (derivado)	شد, شش, شانه, شن
18	تيم	-	تيم, تن, تا
19	هم	también	هم, همه, همين, همت, همتون
20	شود	ser, estar (derivado)	شود

A partir de los datos obtenidos podemos fácilmente constatar que todas estas palabras son palabras vacías en la lengua persa, pero son consideradas como palabras claves por Google (Sadeghi, et al., 2014). Hay una palabra en la fila 18 (-) que no tiene ningún significado y eso ocurre en muchos otros casos en la tabla de contenidos. Esto puede explicarse por el hecho de que la tokenización de las palabras no es correcta y una parte de la palabra original se quedó aislada debido a la presencia de un espacio

blanco entre la palabra. Desde el punto de vista gramática las palabras vacías no tienes palabras derivadas a la excepción del caso de una palabra vacía verbal (la conjugación en diferentes tiempos de un verbo). Pero en nuestro caso, hay muchas variantes encontradas para una palabra vacía y algunas de ellas no tienen ningún significado. Por ejemplo, la palabra en la octava fila را [ra] es el marcador del objeto en la lengua persa con sólo dos letras pero, tiene siete variantes. Una de sus variantes es la palabra راست [rast] (derecha, en español) y otra رامون [ramon] que no tiene ningún significado.

Por lo general, en la tabla de palabras claves hemos encontrado muchas palabras que aunque no son palabras vacías sus variantes son falsas. Por ejemplo la palabra دانش [danesh] (conocimiento, en español) tiene 4 variantes: دانه [daneh] (grano), داند [danad] (sabe), دانند [danand] (saben) y دانیم [danim] (sabemos).

Para conocer mejor la estrategia de Google para indexar los documentos persas hemos decidido hacer el mismo proceso, pero esta vez, con un sitio web de documentos en inglés y español de forma que podemos comparar los resultados obtenidos. Por eso, hemos solicitado a Google para analizar el segundo sitio web que contiene 128 páginas web escritas en español e inglés (http://percomp.info.uva.es). En este caso Google ha indexado todas las páginas y mostramos en la Tabla 4.17 las palabras claves y sus variantes que fueron seleccionadas por Google. La tabla está también ordenada según la importancia de las palabras. Para facilitar al lector la lectura de tabla ponemos sólo las 20 primeras palabras claves.

A partir de la información obtenida podemos realizar algunas observaciones. La primera constatación que podemos concluir es que no hay ninguna palabra vacía como palabra clave. En este caso, las palabras vacías pueden ser las del español o del inglés pero no hay ninguna de ellas, ya sea en inglés o en español en la Tabla. La segunda constatación es que las variantes de la palabra clave son todas correctas. La variante de la palabra puede ser de forma singular o plural (por ejemplo la primera fila, móvil y móviles) o bien de forma derivativa (por ejemplo, en la octava fila; educación, educativos y educativa). Entonces, parece que la lematización de las palabras (inglés e español) es correcta. Finalmente, la última constatación que podemos tener es que la

tokenización de las palabras es también correcta sin ambigüedades porque no hay ninguna palabra que sea incorrecta ortográficamente.

Tabla 4.17: Palabras claves del contenido en "percomp.info.uva.es"

Orden	Palabra clave	Variantes encontradas
1	móvil	móvil, móviles
2	estudio	estudio, estudios
3	hestelo	hestelo
4	investigación	investigación, investigaciones
5	proyectos	proyectos, proyecto
6	encuesta	encuesta
7	hábitos	hábitos
8	educación	educación, educativos, educativa
9	dispositivos	dispositivos, dispositivo
10	movilidad	movilidad
11	app	app, apps
12	android	android
13	uva	uva
14	ios	ios
15	innovación	innovación
16	computación	computación
17	cuestionarios	cuestionarios, cuestionario
18	monitorización	monitorización
19	pervasiva	pervasiva
20	cargas	cargas, carga

Entonces si comparamos las tablas de contenidos de dos sitios web, las conclusiones que podemos deducir son, en primero, Google considera las palabras vacías persas como palabras del contenido de un documento. La no eliminación de las palabras vacías introduce un factor de ruido considerable y por consecuencia el resultado de la búsqueda contiene documentos no relevantes. La segunda conclusión es que la tokenización del texto realizada por Google se basa en la aparición del espacio

blanco en el texto mientras que, como hemos visto en el Capítulo 2 en relación con las características de la lengua persa el proceso de la segmentación del texto persa es más complejo. El problema de la tokenización por Google se ve tanto en la selección de las palabras claves que en la identificación de las variantes de una palabra. La correcta tokenización permite la creación correcta de un índice que facilite el acceso a los documentos que contengan los términos que los representan. La tercera constatación es que la lematización del texto persa que en este caso es la construcción de las variantes de una palabra produce las palabras derivadas que no son gramáticamente correctas.

4.11. Conclusiones

En este parte de nuestro trabajo, hemos presentado el rendimiento del buscador Google, el motor de búsqueda más utilizado entre los usuarios iraníes, para extraer la información relevante de los documentos persas en la Web.

Para lograr nuestros experimentos, hemos construido un sitio Web con todos los documentos del corpus Hamshahri (166.774 páginas Web) que es la única colección estándar construida según las características de TREC. Como todos los documentos no han sido indexados por Google debido a la gran cantidad de documentos, por lo que decidimos repartirlos en varios sitios web para disminuir el tamaño del contenido. Entre estos sitios web hay uno que contiene 36.399 documentos que corresponden a los documentos de Hamshahri entre 1996 y 1997 en el que casi el 93% de documentos están indexados por Google (http://farsidoc1.infor.uva.es). Entonces, elegimos este sitio para desarrollar nuestros experimentos.

Presentamos al buscador Google 100 consultas (las temas del corpus) y comparamos las páginas Web devueltas con los documentos relevantes del juicio de relevancia. Luego las medidas obtenidas de evaluación de la recuperación han sido comparadas con unas medidas de referencia y tenemos las siguientes conclusiones:

A partir de curvas de precisión-exhaustividad obtenidas sobre el conjunto de consultas por el sistema de búsqueda de Google constatamos que, a todos los niveles de

exhaustividad, la precisión de la relevancia de los documentos recuperados por Google es menor que la de la base de referencia. La medida de la precisión media, que es una descripción de la eficacia de un sistema de RI en relación a un conjunto de consultas, es también muy inferior en comparación con la precisión media de la base de referencia. Estos resultados nos revelan que los documentos recuperados por Google ante una consulta en persa no son tan relevantes.

La comparación entre las palabras claves del contenido de nuestro sitio web y la lista de las palabras vacías persas nos indica que las palabras vacías no están eliminadas en el análisis del texto por Google y están consideradas como términos de índice. Podríamos, tal vez, concluir que Google representa los documentos persas utilizando su contenido completo sin discriminar las palabras por su aportación al proceso. Es lo que se denomina representación a texto completo del documento. Efectivamente es la forma más completa de representar un documento, pero implica un coste computacional muy alto para colecciones grandes de documentos. En este caso, cada documento se puede representar por todas sus palabras, tanto nombres, como verbos, adjetivos, adverbios, etc. A pesar de ello, no todos los términos poseen la misma utilidad para describir el contenido de un documento. De hecho, hay términos más importantes que otros, pero no es tarea fácil decidir la importancia de cada término.

Analizando los datos de la tabla de palabras claves del contenido, constatamos que el analizador de Google tiene problemas al nivel de la tokenización y la segmentación correcta de palabras en el texto persa. Hay muchas palabras en la tabla que no existen en el léxico persa debido a una tokenización incorrecta. La correcta segmentación y separación de las palabras son muy importantes en un sistema de RI debido a que estas palabras serán candidatas a ser adoptadas como términos de índice por el sistema.

Las variantes encontradas por Google para las palabras claves no son tampoco correctas. Hay muchas palabras donde las variantes no tienen ninguna relación con la palabra original. Otro problema del analizador de Google es el hecho de encontrar las variantes para las palabras vacías persas. Desde el punto de vista gramática, las palabras

vacías no tienen palabras derivadas a la excepción del caso de una palabra vacía verbal (la conjugación en diferentes tiempos de un verbo). Los algoritmos de Google al nivel de reglas gramaticales de la lengua persa deben ser mejorados para poder construir correctamente las variantes de palabras en el texto.

Finalmente la ultima conclusión que podemos deducir es, cuando analizamos los resultados de precisión para cada consulta separadamente, podemos constatar que los mejores valores de precisión son para aquellos documentos recuperados que tienen nombres propios en las consultas solicitadas.

Capítulo 5

Construcción Automática de Palabras Vacías para Sistemas de Recuperación de Información en Persa

Resumen

La identificación de las palabras vacías es una de las tareas más importantes para muchas aplicaciones de procesamiento del texto tales como la recuperación de información. Las palabras vacías son aquellas palabras que son demasiado frecuentes entre los documentos de un corpus y no contribuyen de manera significativa para determinar el contexto o la información sobre los documentos. La palabra vacía no tiene valor como término de índice y debería ser excluida durante el proceso de indexación antes de hacer consultas por un sistema de recuperación de información. En este capítulo, proponemos un método automático basado en la frecuencia de los términos, la frecuencia inversa normalizada y el modelo de información del documento para extraer las palabras vacías ligeras del texto persa. Denominamos "palabra vacía ligera" a una palabra vacía que tiene muy pocas letras y no es una palabra compuesta. En la lengua persa, una lista completa de palabras vacías se puede deducir combinando las palabras vacías ligeras. Los resultados de la evaluación, usando un corpus estándar, muestran un buen porcentaje de coincidencia entre las palabras vacías de inglés y de persa y también una mejora significativa en el tamaño de los términos de índice. En concreto, las 32 primeras palabras vacías ligeras tienen un gran impacto en la reducción del tamaño de índice y el conjunto de palabras vacías puede reducir el tamaño de los términos de índice hasta un 27%.

5.1. Introducción

Una de las tareas iníciales para construir un sistema de RI consiste en la identificación de las palabras que son demasiado frecuentes entre los documentos de la colección. Por definición, las palabras vacías (*stopword*, en inglés) son palabras muy comunes que aparecen en el texto con frecuencia pero no conllevan la información importante en términos de RI. Este conjunto de palabras se compone de preposiciones, artículos, adverbios, conjunciones, posesivos, demostrativos, pronombres, algunos verbos (como

ser, estar en español) y algunos nombres. Las palabras vacías tienen un impacto significativo en el proceso de recuperación de textos en diferentes idiomas. Mediante la exclusión de estas palabras se reduce el tamaño del índice y generalmente mejora la eficacia de recuperación (Korfhage, 1997). En TREC, las 33 principales palabras vacías representan el 30% de todas las palabras (Witten, et al., 1999). De acuerdo con Kucera (Kucera, et al., 1982) las diez palabras más frecuentes en inglés generalmente representan del 20 al 30% de *tokens* en un documento. Estas palabras tienen un valor muy bajo de discriminación porque el resultado será un conjunto de documentos irrelevantes cuando se consideran para el propósito de la búsqueda (Van-Rijsbergen, 1979). Es decir, la cantidad de información transportada por estas palabras no es significante. En consecuencia, es útil generalmente eliminar todas estas palabras al indexar los documentos y al procesar las consultas.

Una forma de mejorar el rendimiento de un sistema de RI es entonces, la eliminación de palabras vacías en la fase de indexación automática. Tradicionalmente, se supone que la lista de palabras vacías contiene las palabras más frecuentes que ocurren en un documento. Sin embargo, en el análisis léxico del texto aparece que algunas palabras con frecuencia en el texto son también importantes como términos del índice. En general, la construcción de la lista de palabras vacías dependerá de documentos de la base de datos, características del usuario y el proceso de indexación. El uso de una única lista de palabras vacías a través de diferentes colecciones de documentos podría ser perjudicial para la eficacia de recuperación.

Muchas de las listas de palabras vacías se han desarrollado para el idioma inglés y se basan generalmente en las estadísticas de frecuencia de un gran corpus. Las listas de palabras vacías en inglés disponibles en la Web son buenos ejemplos de ellas (WordNet, 2007) (XPO6, 2013). Pero la investigación y la experimentación en el campo de la RI en la lengua persa son relativamente nuevas y limitadas en comparación con la investigación que se ha hecho en otros idiomas, sobre todo en inglés. Uno de los problemas encontrados en la RI con la lengua persa es el de las palabras vacías. Como hemos descrito en la Sección 0 las listas de palabras vacías persas disponibles son casi todas derivadas de pequeñas colecciones de textos basándose en la alta frecuencia de los términos en la colección. Estas listas son relativamente cortas e incompletas y

fueron editadas manualmente para añadir o eliminar algunas palabras. En nuestro mejor entendimiento, podemos afirmar que no hay ningún trabajo de investigación que consista en identificar automáticamente las palabras vacías en un sistema de RI en persa.

En efecto, parece que la construcción de listas de palabras vacías en persa representa un escollo importante; una razón para ello es, desde luego, el desconocimiento de la lengua por parte de los investigadores, no todos ellos farsi parlantes. Pero otro es la carencia (o su desconocimiento, al menos) de corpus y estudios estadísticos para el idioma persa. En contra de lo que pudiera pensarse, la eliminación o no de palabras vacías no es simplemente una cuestión de tamaños de índices y ficheros invertidos (y de tiempo de procesamiento); más aún, con el precio que la memoria tiene actualmente, éste sería un problema nimio. El problema radica en que, con sistemas que atribuyen pesos a términos, y que operan con éstos, las palabras vacías introducen un factor de ruido considerable. En el capitulo anterior, hemos visto este fenómeno por el buscador Google frente a los documentos persas. El hecho de que Google considere palabras vacías como palabras claves en un documento persa, hace que los usuarios se enfrenten a documentos recuperados que no son tan relevantes.

Uno de los problemas y desafíos en el procesamiento del texto persa es las diferentes formas de escritura. Otra dificultad que pueda afectar a la identificación automática de las palabras vacías es cómo determinar el límite de la palabra. Como hemos mencionado anteriormente en la Sección 2.6.3.1, el espacio blanco dentro de un texto persa no es el signo de límite y un separador entre las palabras. Puede aparecer dentro de una palabra o entre palabras. Por otro lado, puede que no haya espacio entre dos palabras. Hay muchas palabras que pueden ser escritas con el espacio, el espacio corto o sin ningún espacio. Es por eso que la Academia de la Lengua y de la Literatura Persa (*APLL*)[28] introdujo las nuevas reglas y estilos estándares. Una de las recomendaciones de *APLL* es escribir por separado los adverbios, preposiciones y conjunciones en las oraciones. Generalmente, estas categorías de palabras son menos

[28] Academy of Persian Language and Literature

importantes y pueden ser considerados como palabras vacías. Por ejemplo, la palabra "یک" [yek] (uno, en español) y "هیچ" [hich] (cualquier, en español) son considerados como palabras vacías. Nosotros llamamos estas palabras las "palabras vacías ligeras", porque son simples (no compuestas) y también tienen muy pocas letras. La palabra compuesta con las dos palabras es "هیچ یک" [hichyek] (ninguno, en español). La palabra compuesta obtenida es también una palabra vacía y esto ocurre en muchos otros casos en la lengua persa. Entonces, nuestro objetivo es identificar automáticamente las palabras vacías ligeras y por la combinación de ellas, obtener una lista completa de las palabras vacías. Nuestro método está basado en los modelos estadísticos y en el modelo de información. El modelo estadístico extrae las palabras vacías teniendo en cuenta la distribución de estas palabras en un corpus y en cada documento del corpus. El modelo de información mide el significado de una palabra en el texto mediante el uso de la teoría de la información. Los resultados de estos dos modelos están agregados para generar una lista de palabras vacías para sistemas de RI textuales en la lengua persa.

5.2. Trabajos Relacionados

Por lo general, los documentos son los objetos principales en un sistema de RI y un cierto proceso se lleva a cabo en ellos para que el sistema sea listo para funcionar. Uno de los procesamientos léxicos tradicionales incluye la construcción de la lista de palabras vacías. Algunas investigaciones relacionadas han sido desarrolladas para el idioma inglés. Por ejemplo, Francis y Kucera (Kucera, et al., 1982) trabajaron en el Brown Corpus y fueron capaces de extraer 425 palabras vacías. Del mismo modo, Van-Rijsbergen (Van-Rijsbergen, 1979) elaboró una lista de palabras vacías para el inglés que comprende 250 palabras de alta frecuencia y palabras "*fluff*". La palabra "*fluff*" es, por ejemplo como *below*, *near* o *alway*s que tiene frecuencia bastante baja en un texto inglés pero semánticamente no tiene un poder de discriminación significativa.

Tsz-Wai Lo (Tsz-Wai Lo, et al., 2005) propuso un nuevo método, llamado "*term-based random sampling*" para generar automáticamente una lista de palabras vacías para una dada colección. Este enfoque, inspirado en la técnica de expansión de

consultas, se basa en cómo de informativo es un término dado. La importancia de un término puede evaluarse utilizando la medida de la divergencia de Kullback-Leibler (Cover, et al., 2006). Este enfoque se compara entonces con diversos enfoques clásicos basados en la ley de Zipf considerando como métodos de líneas de base. Los resultados muestran que las listas de palabras de vacías derivadas de los métodos inspirados en la ley de Zipf son fiables, pero muy costoso para llevar a cabo. Por otro lado, el esfuerzo computacional necesario para obtener las listas de palabras vacías utilizando el nuevo enfoque fue mínimo en comparación con los enfoques clásicos. Por último, se muestra que una lista de palabras vacías más eficaz se puede conseguir mediante la fusión de la lista de palabras vacías clásica con las listas de palabras vacías generadas por cualquiera de las líneas de base o el nuevo enfoque propuesto.

Feng Zou (Zou, et al., 2006) sugirió un método para elaborar una lista de palabras vacías para el idioma chino. Se utiliza un modelo agregado para medir tanto la característica de la frecuencia de una palabra por modelo estadístico y su característica de información por el modelo de información. Este enfoque ha sido desarrollado basado en la idea de que las palabras vacías están ordenadas en la parte superior con una frecuencia mucho mayor que las otras palabras, mientras que al mismo tiempo, mantienen una distribución estable en diferentes documentos. Una combinación de estas dos observaciones redefine las palabras vacías como palabras con frecuencias estables y altas en los documentos. La lista generada se comparó con otras listas existentes y mostró una mejora con respecto a los otros enfoques. Alajmi (Alajmi, et al., 2012) utiliza el mismo método para generar una lista de palabras vacías para la lengua árabe.

El-khair (El-Khair, 2006) llevó a cabo un estudio comparativo sobre el efecto de la eliminación de palabras vacías en la RI en árabe. Tres listas de palabras vacías fueron utilizadas para hacer una comparación que son palabras vacías generales, basadas en el corpus y combinadas. La lista general fue creada basándose en las características de estructura de la lengua árabe. La segunda se elaboró basada en la frecuencia de palabras en el corpus y la tercera fue creada combinando las dos. Se concluyó que la combinación de las listas y la lista general de palabras vacáis producen las mejores funciones de rendimiento para la recuperación en la lengua árabe usando el

algoritmo *BM25*. El rendimiento de la lista general y la lista combinada fue relativamente cercano. Se recomienda el uso de cualquiera de ellos, pero la lista general es ciertamente preferible cuando se trata de aplicar para diferentes corpus.

En cuanto a la lengua persa, la Sección 0 nos da un detalle de los trabajos previos en este ámbito. Como hemos visto sólo hay unos pocos estudios sobre la construcción de palabras vacías persas. La lista presentada por Taghva (Taghva, et al., 2003b) contiene 155 y 12 palabras vacías verbales. Las listas construidas que contienen las palabras de alta frecuencia en el corpus Hamshahri (AleAhmad, et al., 2009) y Mahak (Esmaili, et al., 2007) son relativamente cortas e incompletas y sus palabras son todas incluidas en la lista de Taghva (Taghva, et al., 2003b). La última lista está identificada por Davarpanah (Davarpanah, et al., 2009) conteniendo 927 términos. Su construcción fue manual por expertos combinando las palabras para crear nuevas palabras y extender la lista.

5.3. Construcción de Palabras Vacías Ligeras

Una revisión lingüística y gramática de la lengua persa (Davarpanah, et al., 2009) (Safavi, 1981) (Bateni, 2003) revela que las palabras de la lengua persa, como otras lenguas, tienen dos niveles distintos de representación: la representación semántica y la sintáctica. Como se menciona en la literatura relacionada, las palabras vacías tienen principalmente una función sintáctica. De hecho, se utilizan sólo para la construcción gramatical de oraciones y no llevan ninguna información significativa (Zou, et al., 2006). Por lo tanto, las posibles palabras que pueden ser consideradas como palabras vacías deben ser recogidas de las diferentes clases sintácticas de una manera sistemática para asegurar la integridad de la lista. Desde el punto de vista de la lingüística, las palabras vacías persas generalmente serán palabras en las siguientes categorías de palabras: adverbios de tiempo y lugar, pronombres, preposiciones, conjunciones, determinantes, interjección, palabras interrogativas, números ordinales, auxiliares y algunos verbos (palabras vacías verbales). En general, la palabra vacía de la lengua persa, como en otros idiomas, tiene ciertas propiedades:

- tiene poco sentido si se utilizan por separado;
- aparece muchas veces en un texto;
- son necesarios para la construcción de sentencias;
- es una palabra general y no se utiliza, sobre todo, en un campo determinado;
- no se utiliza como palabra clave de búsqueda;
- nunca forma una oración completa cuando se utiliza aisladamente.

El método de construcción de palabras vacías ligeras tiene varios pasos. Nuestra hipótesis es considerar que una palabra vacía ligera persa tiene una longitud pequeña (muy pocas letras) y un conjunto completo de palabras vacías puede derivarse mediante la combinación de ellas. Con el fin de determinar la lista de palabras vacías ligeras, seguiremos una metodología basada en agregación. La primera etapa será la identificación de una lista de palabras de uso frecuente en el léxico persa. El segundo paso será la generación de una lista de términos que tienen un valor bajo de frecuencia inversa del documento. En el último paso, calculamos la medida de la entropía para cada palabra y construimos una lista de palabras con valor alto de entropía. Por último, las tres listas obtenidas se agregarán para derivar una lista final.

A continuación de nuestro trabajo, vamos a utilizar la versión original del corpus Hamshahri (ver Sección 3.2.4). Como hemos indicado previamente, Hamshahri es una colección de prueba estándar y la más grande para el texto persa y está construida según las especificas de TREC, por lo que resulta ser la más adecuada. Los documentos de esta versión están en formato texto y podemos fácilmente extraer los términos y calcular sus frecuencias. Hay un total de 166.774 documentos que contienen más de 63 millones de palabras con 417.339 palabras únicas.

5.3.1. Longitud de Palabras

Las palabras con alta frecuencia en el texto persa tienen una característica muy interesante. No son parte de palabras compuestas y, en general, su longitud (número de caracteres) varía de 2 a 5 letras. Un estudio estadístico de las propiedades de palabras

pesas nos mostró esta característica. La Figura 5.1 muestra la ocurrencia total de palabras de n-letras (*token*) en el corpus Hamshahri. Como podemos ver, las palabras entre dos y cinco letras, tienen una alta frecuencia en la colección. En consecuencia, las palabras de n-letras (n = 2,..., 5) son la base de nuestros experimentos para identificar las palabras vacías ligeras.

De las palabras únicas (417.339 palabras), primero extraemos todas las palabras entre 2 a 5 letras (79.989 palabras en el corpus). Hay sólo una palabra con una sola letra que tiene significado en el texto persa. Esta palabra es "و" (y o e, en español) y tiene una frecuencia muy alta en relación con otras palabras de la colección (aproximadamente 2.850.000 ocurrencias). Esta palabra ocupa el primer lugar de la lista de palabras vacías con todos los métodos que vamos a presentar en detalle.

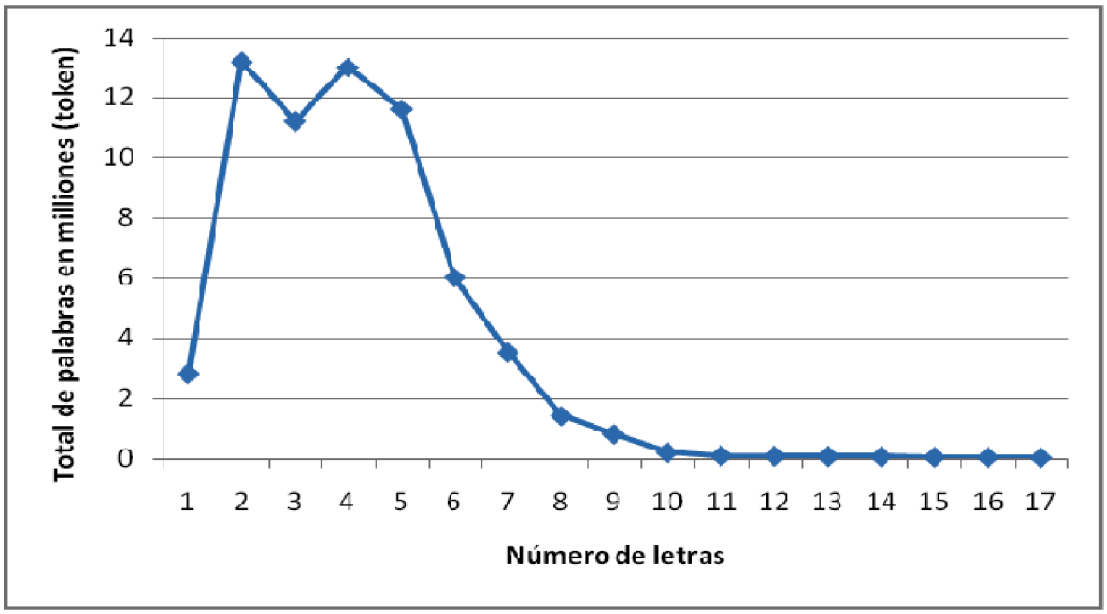

Figura 5.1: El número total de palabras de n-letras (*tokens*) en el corpus Hamshahri

5.3.2. Frecuencia de los Términos en la Colección

Una de las características más evidentes del texto desde el punto de vista estadístico es que la distribución de frecuencias de palabras es muy sesgada. Es decir que unas pocas palabras (palabras vacías) aparecen muchas veces y muchas aparecen pocas veces en

una colección de textos documentos. George Kingsley Zipf (1902-1950) observó que la distribución del orden de la frecuencia de los términos es muy cerca de la relación:

$$F(r) = \frac{C}{r^{\alpha}} \qquad (5.1)$$

donde $\alpha \approx 1$ y $C \approx 0.1$. La ecuación 5.1 se conoce como ley de Zipf (Zipf, 1949) y afirma que la frecuencia de cualquier palabra es inversamente proporcional a su posición en la tabla de frecuencias. Por consecuencia, una lista de palabras vacías se puede deducir simplemente usando la *n* palabras más frecuentes en una colección de documentos. El hecho de que la naturaleza y las características de la lengua persa son diferentes de las de inglés antes, verificamos si bien la ley de Zipf puede ser válida para la lengua Persa.

5.3.2.1. Verificación de la Ley de Zipf en el Texto Persa

Para tener el número de ocurrencia de cada palabra en un documento, hemos construido un sencillo software que calcula la ocurrencia total de una palabra en un documento y después nos da una tabla ordenada de palabras según su frecuencia de aparición en el texto.

Tabla 5.1: Una parte de la distribución de los términos en un texto persa

Rango	Palabra	Frecuencia	Porcentaje de palabra en el documento	En español
1	و	209	4,88	y, e
2	در	123	2,87	en
3	که	117	2,73	que
4	به	87	2,03	a
5	دانش	84	1,96	conocimiento
6	از	81	1,89	desde
7	است	64	1,50	es
8	یک	57	1,33	uno
9	مدیریت	57	1,33	gestión
10	این	49	1,14	esto, eso

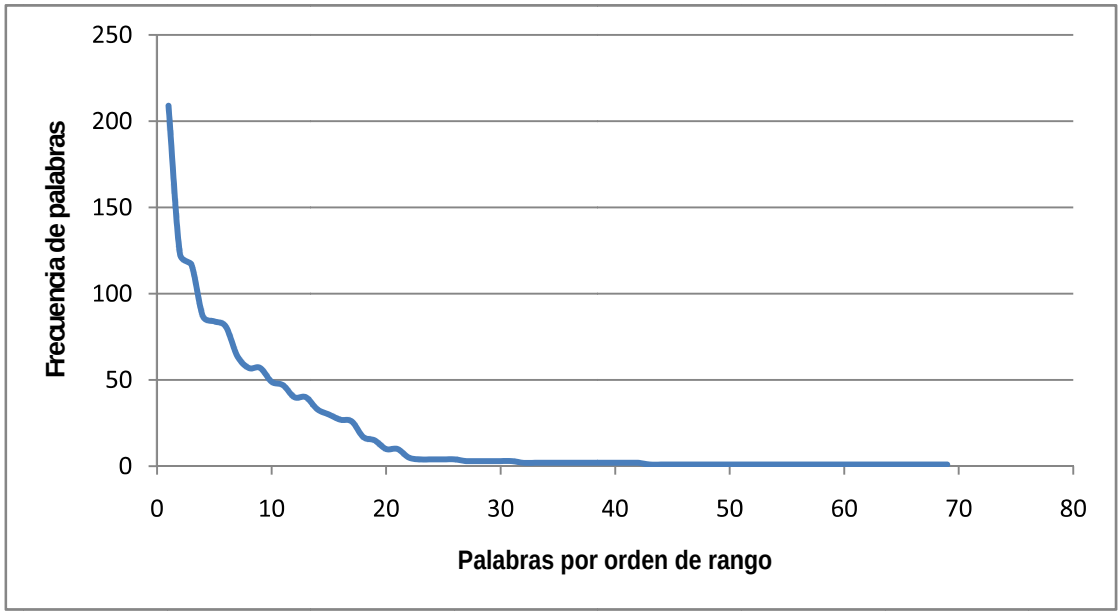

Figura 5.2: La ley de Zipf en un texto persa

Figura 5.3: Ley de Zipf en la colección Hamshahri

En primero, aplicamos nuestro método a un texto que está en la parte de apéndice (ver Apéndice C). La Tabla 5.1 muestra las 10 primeras palabras más

frecuentes en el mismo texto. La Figura 5.2 muestra el diagrama de la curva relacionado la frecuencia de la ocurrencia y la orden de palabras en el mismo documento. En la segunda etapa, aplicamos el mismo proceso a todos los documentos de la colección Hamshahri y tenemos la curva obtenida en la Figura 5.3 a partir del conjunto de todas las palabras del corpus. En este diagrama la frecuencia de palabras tiene una escala logarítmica. Como podemos constatar las dos curvas obtenidas parecen a la curva hiperbólica descrita por la ecuación 5.1 y por secuencia podemos identificar una lista de palabras vacías teniendo en cuenta la alta frecuencia de las palabras en la colección.

Tabla 5.2: Las 20 primeras palabras persas con frecuencias más altas

Rango	Palabra	Longitud	En español
1	و	1	y, e
2	در	2	en
3	به	2	a
4	از	2	desde
5	که	2	que
6	این	3	esto, este
7	می	2	partícula verbal del imperativo
8	است	3	es
9	را	2	râ (marcador del objeto)
10	با	2	con
11	های	3	sufijo usado como plural + *ezafe*
12	برای	4	para, por
13	آن	2	eso
14	ها	2	sufijo usado como plural
15	يك	2	uno
16	شود	3	estar, ser (derivado)
17	شده	3	estar, ser (derivado)
18	خود	3	si mismo
19	كرد	3	hizo
20	ای	2	marcador del indefinido

La Tabla 5.2 muestra la lista de las 20 primeras palabras con mayor frecuencia y está ordenada en orden descendente de acuerdo con la frecuencia de los términos.

5.3.3. Frecuencia de los Términos en el Documento

Luhn (Luhn, 1957) utilizó la ley de Zipf como hipótesis nula para permitirle especificar dos umbrales, *upper cut-off* y *lower cut-off* (denotados como corte superior y corte inferior en la figura de la ecuación de Zipf), tratando de excluir así a las palabras no significativas. Los términos que excedían en frecuencia el corte superior eran considerados palabras de uso común, mientras que las que no llegaban al corte inferior se consideraban poco comunes, y no contribuirían perceptiblemente a describir el contenido del documento. Las palabras que excedan el límite superior se puede considerar como palabras vacías, ya que tienen un bajo peso en el proceso de indexación.

En otras palabras, los términos que aparecen con poca frecuencia tienen una mayor probabilidad de ocurrir en los documentos apropiados y deben ser considerados como más informativos y por lo tanto de más importancia en estos documentos. Mediante la sustitución de "frecuencia de los términos" con "frecuencia inversa de documentos normalizada" (*IDF, Inverse Document Frequency*), se puede construir otra lista de palabras vacías. *IDF* es el coeficiente que determina la capacidad discriminatoria del término de un documento con respecto a la colección. Es decir, distinguir la homogeneidad o heterogeneidad del documento a través de sus términos. Si un término tiene un poder discriminatorio bajo, entonces ese término es genérico y aparece en la mayoría de los documentos. *IDF* normalizada es la forma más común de ponderación de *IDF*, utilizado por Robertson y Sparck-Jones (Robertson, et al., 1976), que se normaliza con respecto al número de documentos que no contengan el término (N_{doc}– D_k) y añadiendo una constante de 0.5 tanto al numerador que al denominador para moderar los valores extremos:

$$idf_k = \log_2\left(\frac{(\mathrm{N_{doc}} - \mathrm{D_k}) + 0.5}{\mathrm{D_k} + 0.5}\right) \qquad (5.2)$$

N_{doc} es el número total de documentos en la colección y D_k es el número de documentos que contienen el término k. La palabra vacía es aquella haciendo un término de índice muy pobre y naturalmente tiene un bajo peso de frecuencia inversa de documento. Calculamos la frecuencia inversa de documento para cada término de la colección. El valor de idf_k varía de -4,48958 hasta +16,76257. Las 18 primeras palabras con menor valor de idf_k tienen valores negativos. El valor máximo de idf_k corresponde a las palabras que se encuentran en un único documento de la colección. La Tabla 5.3 muestra los 20 principales palabras persas con valor bajo de idf_k y están ordenadas de forma ascendente al valor idf_k.

Tabla 5.3: Las 20 primeras palabras con ponderación más baja de *IDF*

Rango	Palabra	Valor idf_k	Longitud	Traducción español
1	و	-4,48958	1	y, e
2	در	-4,09851	2	en
3	به	-3,78071	2	a
4	از	-3,17551	2	desde
5	این	-2,68658	3	esto, este
6	با	-2,15464	2	con
7	است	-1,87941	3	es
8	که	-1,87658	2	que
9	می	-1,84175	2	partícula verbal del imperativo
10	را	-1,61037	2	râ (marcador del objeto)
11	های	-1,22565	3	sufijo usado como plural + *ezafe*
12	برای	-0,87331	4	para, por
13	شد	-0,42322	2	ser, estar (derivado)
14	شده	-0,42014	3	ser, estar (derivado)
15	شود	-0,41371	3	ser, estar (derivado)
16	کرد	-0,40513	3	hizo
17	يك	-0,19271	2	uno
18	آن	-0,01135	2	eso
19	خود	0,02740	3	si mismo
20	تا	0,20839	2	hasta

5.3.4. Modelo de Información

La entropía es una medida de imprevisibilidad o el contenido de información (Ribeiro, 2004). La entropía también se puede considerar como la cantidad de información promedio que contienen los símbolos usados. Los símbolos con menor probabilidad son los que aportan mayor información; por ejemplo, si se considera como sistema de símbolos a las palabras en un texto, palabras frecuentes como "que", "el", "a" aportan poca información, mientras que palabras menos frecuentes como "corren", "niño", "perro" aportan más información. Si de un texto dado eliminamos un "que", seguramente no afectará a la comprensión y se sobreentenderá, no siendo así si borramos la palabra "niño" del mismo texto original. Cuando todos los símbolos son igualmente probables (distribución de probabilidad plana), todos aportan información relevante y la entropía es máxima.

La entropía de Shannon, definida por Claude E. Shannon (Shannon, 1948), es una función matemática que corresponde intuitivamente a la cantidad de información contenida o suministrada por una fuente de información. Esta fuente puede ser texto escrito en un idioma particular o cualquier archivo de computadora (colección de bytes). Desde el punto de vista de la teoría de la información, las palabras vacías son también aquellas palabras que llevan poca información. La entropía, una de las medidas fundamentales de información(Zou, et al., 2006), nos ofrece otro método para describir mejor la selección de palabras no significativas.

Supongamos que hay M palabras distintas y N documentos. Denotamos cada palabra como w_j (j = 1,..., M) y cada documento como D_i (i = 1,..., N). Para cada palabra w_j, se calcula su frecuencia en el documento D_i denotada como $f_{i,j}$. Sin embargo, el documento tiene diferente longitud. Con el fin de normalizar la longitud del documento, se calcula la probabilidad $P_{i,j}$ de la palabra w_j en el documento D_i que es su frecuencia en el documento D_i dividido por el número total de palabras en el documento D_i. Por lo tanto, se mide el valor de la información de la palabra w_j por su entropía. Calculamos el valor de la entropía (H) para la palabra w_j de la siguiente manera:

$$H(w_j) = \sum_{i=1}^{N} P_{i,j} \times \log_2\left(\frac{1}{P_{i,j}}\right) \qquad (5.3)$$

Una vez que se ha calculado la entropía de cada palabra en el conjunto de datos, la lista resultante se puede ordenar por el ascendente de entropía para revelar las palabras que tienen una mayor probabilidad de ser las palabras irrelevantes o ruidos. Por medida de la entropía de cada palabra, podemos preparar otra lista ordenada de palabras.

Tabla 5.4: Las 20 primeras palabras persas con mayor valor de entropía

Rango	Palabra	Entropía	Longitud	En español
1	و	0,204825	1	y, e
2	در	0,170488	2	en
3	به	0,158113	2	a
4	از	0,134870	2	desde
5	که	0,129905	2	que
6	می	0,120336	2	partícula verbal del imperativo
7	این	0,113495	3	esto, este
8	را	0,108043	2	râ (marcador del objeto)
9	است	0,105118	3	es
10	با	0,088965	2	con
11	های	0,088603	3	sufijo usado como plural + *ezafe*
12	ها	0,072121	2	sufijo usado como plural
13	آن	0,068665	2	eso
14	برای	0,064404	4	para, por
15	يك	0,059400	2	uno
16	خود	0,056542	3	si mismo
17	سال	0,056464	3	año
18	شود	0,055342	3	ser, estar (derivado)
19	بود	0,054736	3	fue
20	شده	0,054631	3	ser, estar (derivado)

La palabra que tiene el valor más alto de entropía es la que posee el valor inferior de información(Zou, et al., 2006). Por lo tanto, las palabras con mayor valor de entropía se extraen como candidatos para las palabras vacías. En la Tabla 5.4 se muestran las 20 primeras palabras con el valor de entropía más alto.

5.3.5. Agregación

Las características de las palabras vacías se revelan en diferentes aspectos generados por las tres listas ordenadas. ¿Cómo conseguir una agregación de ellas? ¿Qué tipo de reglas podría asegurar la equidad del resultado final? Una de las soluciones populares a esto debería ser la regla de Borda[29] (Myerson, 2013). El recuento de Borda es un método de elección de un solo ganador en el cual los votantes clasifican los candidatos por orden de preferencia. El recuento de Borda determina el ganador de una elección, dando a cada candidato un número determinado de puntos que corresponden a la posición en la que él o ella están en el puesto por cada votante. Una vez que todos los votos se han contado, el candidato con más puntos es el ganador. El número de puntos otorgados a los candidatos para cada clasificación se determina por el número de candidatos que se presentan en las elecciones.

La regla de Borda es un método de votación en la teoría de la elección social mediante el que los votantes manifiestan sus preferencias a través de órdenes lineales sobre los candidatos, asignando a éstos puntuaciones escalonadas según su mérito (Myerson, 2013). Utilizando la terminología de la literatura de votación, podemos ver cada palabra vacías de una lista ordenada como candidato y cada lista obtenida como un votante. Cada candidato recibe puntos de cada uno de los votantes, de acuerdo con su rango en la lista de votantes.

Por ejemplo, el candidato mejor clasificado recibirá n puntos, donde n es el número de candidatos en la lista de clasificación respectiva. La puntuación total de

[29] El recuento de Borda fue desarrollado independientemente varias veces, pero se llama así por el matemático y politólogo francés del siglo 18 Jean-Charles de Borda, quien ideó el sistema en 1770.

Borda del candidato será la suma de las puntuaciones por cada lista clasificada en el que aparece. En caso de que el candidato no se encuentra en la lista *top-k* de algún elector recibirá una parte de los puntos restantes del votante (cada elector tiene un número fijo de puntos disponibles para su distribución). Usando la regla de Borda, obtenemos las 20 primeras palabras vacías ligeras en persa que están clasificadas en la Tabla 5.5.

Tabla 5.5: Las 20 primeras palabras vacías persas aplicando la clasificación de Borda

Rango	Palabra	Longitud	Traducción español
1	و	1	y, e
2	در	2	en
3	به	2	a
4	از	2	desde
5	این	3	esto, este
6	که	2	que
7	می	2	partícula verbal del imperativo
8	است	3	es
9	با	2	con
10	را	2	râ (marcador del objeto)
11	های	3	sufijo usado como plural + *ezafe*
12	برای	4	para, por
13	آن	2	eso
14	يك	2	uno
15	ها	2	sufijo usado como plural
16	شود	3	ser, estar (derivado)
17	شده	3	ser, estar (derivado)
18	خود	3	si mismo
19	كرد	3	hizo
20	ای	2	marcador del indefinido

5.4. Análisis de los Resultados

Desde el punto de vista lingüística, las palabras vacías persas, similar a las palabras vacías de español, suelen ser esas palabras con parte de oración como adverbios, preposiciones, interjecciones y auxiliares. De acuerdo con diferentes dominios, podríamos clasificar todas las palabras vacías en dos categorías. Un tipo denominado “dominio independiente” o "palabras vacías genéricas" que son palabras vacías en el dominio general. Otro tipo es las palabras vacías “dependientes del dominio” o del documento denominado "palabras vacías del dominio". Como hemos aplicado nuestro método a las palabras con n-letras (n = 2,..., 5), se obtiene en la mayor parte las palabras vacías genéricas, pero también encontramos muy pocas palabras que dependen del dominio como por ejemplo "کشور" [keshvar] (país, en español) con 4 letras.

La comparación de nuestros resultados con la lista de palabras vacías persa ya disponible se muestra en la Tabla 5.6. Nuestra lista se comprobó contra la lista de palabras vacías identificada por Davarpanah (Davarpanah, et al., 2009). La dicha lista está basada en el diccionario y la opinión de los expertos. Los autores extienden su lista mediante la combinación de las palabras vacías que ya existen en la lista y, por lo menos semánticamente, esta lista puede tener un gran potencial para realizar una comparación. La comparación muestra que existe un alto porcentaje de coincidencia. La diferencia de coincidencia entre estas dos listas se explica por el hecho de que el corpus era diferente para la generación de la lista de palabras vacías. Mientras tanto, otra diferencia de estas dos listas se produce en su método de construcción.

Tabla 5.6: Comparación de superposición de palabras vacías persa y las de genéricas en inglés

Nº de palabras vacías en la parte superior de la lista	Superposición con palabras vacías inglés	Coincidencia con palabras vacías persas
10	90%	100%
20	80%	100%
50	82%	90%
100	81%	89%

Otra comparación fue realizada entre la lista generada por nuestro algoritmo y la lista de palabras vacías del corpus Brown (Kucera, et al., 1982) que es muy conocida y ampliamente utilizada en inglés (ver Tabla 5.6). Los resultados muestran una alta coincidencia entre las palabras vacías genéricas en inglés y las 100 primeras palabras vacías persas. Esta similitud es más fuerte en las 10 primeras palabras vacías persas. En este caso particular, hay solo una palabra, que es el signo de partícula verbal imperfectivo en lengua persa, sin equivalente en inglés. La diferencia de coincidencia se explica por el hecho de que, en general, un gran número de palabras vacías se deben a las características del lenguaje. Estos dos idiomas son diferentes en su naturaleza y, sin duda, hay algunas palabras sin equivalencia entre ellos.

Como se mencionó anteriormente, el hecho de eliminar las palabras vacías en un sistema de RI reduce también el tamaño del índice. En consecuencia, la eliminación de las palabras vacías disminuye el tiempo de búsqueda para una consulta determinada y no tendría ninguna influencia en la eficacia de recuperación. En la Figura 5.4 se puede observar que las 20 principales palabras vacías ligeras persas reducen alrededor de 22% el tamaño del índice del texto persa.

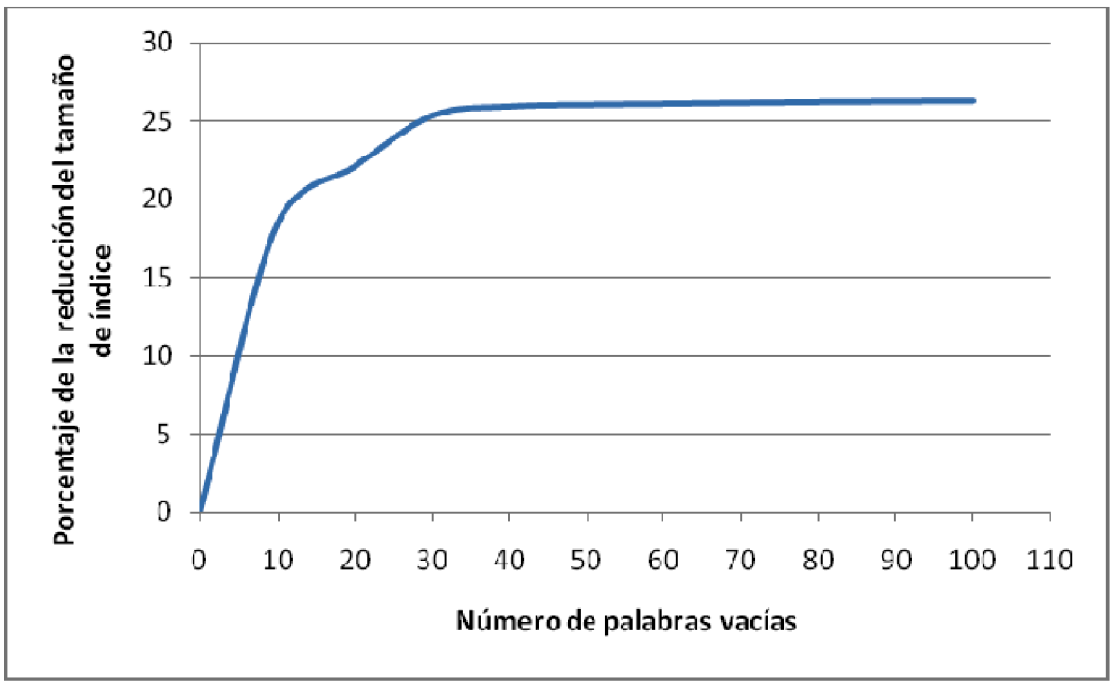

Figura 5.4: Contribución de palabras vacías a la reducción del tamaño de índice

Observamos también que las 32 primeras palabras vacías ligeras tienen un gran impacto en la reducción del tamaño de índice. Después de que, al aumentar el número de palabras vacías en el texto, la reducción del tamaño de índice es muy baja. Los resultados experimentales (ver la Figura 5.4) indican que hay un límite en la reducción del tamaño de índice cuando se consideran las 32 primeras palabras vacías. Podemos, por lo tanto, especificar que hay un límite máximo en la contribución de palabras vacías a la reducción del tamaño del índice. El conjunto de palabras vacías persas reduce el tamaño de los términos de índice en un 27% y el número óptimo de las palabras vacías a considerar es 32. La Tabla 5.7 muestra el resto de las palabras vacías ligeras y, junto con las que se muestran en la Tabla 5.5, completa la lista de las 32 primeras palabras vacías persas.

Tabla 5.7: El resto de las 32 primeras palabras vacías (ver la Tabla 5.5)

Rango	Palabra	Longitud	Traducción español
1	و	1	y, e
2	در	2	en
3	به	2	a
4	از	2	desde
5	این	3	esto, este
6	که	2	que
7	می	2	partícula verbal del imperativo
8	است	3	es
9	با	2	con
10	را	2	râ (marcador del objeto)
11	های	3	sufijo usado como plural + *ezafe*
12	برای	4	para, por
13	آن	2	eso
14	يك	2	uno
15	ها	2	sufijo usado como plural
16	شود	3	ser, estar (derivado)
17	شده	3	ser, estar (derivado)
18	خود	3	si mismo
19	كرد	3	hizo

Tabla 5.7 (continuación)

Rango	Palabra	Longitud	Traducción español
20	ای	2	marcador del indefinido
21	شد	2	ser, estar (derivado)
22	تا	2	hasta
23	بود	3	fue
24	کند	3	hacer (derivado)
25	نیز	3	también
26	گفت	3	dijo
27	دارد	4	tiene
28	دیگر	4	otro, otra
29	بر	2	en, al
30	باید	4	debe
31	هر	2	cada
32	نمی	3	partícula negativa verbal del imperativo

Las diez palabras vacías persas más comunes están, en promedio, en el 86% de los documentos persas y además, aproximadamente el 18 por ciento de las palabras de un documento se encuentran entre ellas. Los detalles de las diez palabras más comunes se muestran en la Tabla 5.8, incluyendo el porcentaje de documentos en el que aparece cada una de ellas. La palabra "و" [va] (y o e, en español) es la palabra más usada en un texto persa y ocupa entre el 3,5 y el 4 por ciento de las palabras en el texto y más de 96% de los documentos textuales contienen esa palabra. Mientras la palabra más común en el idioma inglés es la palabra "*the*" con el primer puesto en el texto inglés (World-english, 2012).

Esta diferencia puede deberse al hecho de que no hay ningún artículo definido en oraciones persas mientras que la mayoría de los sustantivos en inglés aparece con el artículo definido. En el texto persa hay la palabra "را" como un marcador del objeto que desempeña el papel del artículo definido. Esta palabra también se encuentra entre las diez palabras vacías más comunes y ocupa la décima posición (ver la Tabla 5.8).

Tabla 5.8: Las 10 palabras más comunes en el texto persa

Rango	Palabra	Longitud	Porcentaje de documentos	Traducción español
1	و	1	96%	y, e
2	در	2	95%	en
3	به	2	93%	a
4	از	2	90%	desde
5	این	3	87%	esto, este
6	که	2	79%	que
7	می	2	78%	partícula verbal del imperativo
8	است	3	79%	es
8	با	2	82%	con
10	را	2	75%	râ (marcador del objeto)

Hay una pregunta que puede plantearse. ¿Cuál es la lista más cercana de la lista definitiva? Para responder a esta pregunta, tenemos que medir la similitud entre cada una de las tres listas y la lista agregada final utilizando el índice de Jaccard. El índice de Jaccard[30], también conocido como el coeficiente de similitud de Jaccard, es una estadística que se usa para comparar la similitud y la diversidad o distancia entre dos conjuntos de muestreo. Este índice solo utiliza los datos de presencia-ausencia. Supongamos que *A* y *B* son dos conjuntos finitos no vacíos. El coeficiente de Jaccard se define como el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestreo:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (5.4)$$

[30] Paul Jaccard (18 de noviembre 1868 , Sainte-Croix - 9 de mayo 1944 , Zurich) fue un ecólogo, profesor de botánica, fisiólogo vegetal de la Escuela Politécnica Federal de Zúrich. Estudió en la Universidad de Lausana y en la ETH Zurich, obteniendo su PhD en 1894. Continuó estudios en París con Gaston Bonnier. Desarrolló el índice Jaccard de similitud (que llamó coeficiente de comunidad) y publicó en 1901.

Si *A* y *B* son dos conjuntos ambos vacíos, definimos *J(A,B)* = 1. Claramente tenemos $0 \leq J(A,B) \leq 1$. En la Tabla 5.9, se muestra el índice de Jaccard para cada lista de las 20 primeras palabras vacías ligeras con respecto a la lista final. La mayor similitud se da entre la lista definitiva y la lista de términos con alta frecuencia en la colección. En este caso, el valor del índice de Jaccard es igual a uno. En otras palabras, todos los elementos se cruzan entre las dos listas. Además, se ha calculado la distancia entre cada lista y la lista final. La distancia de *Spearman*[31] *Footrule* es una distancia absoluta entre dos vectores ordenados. Esta medida nos da el desplazamiento o la permutación total de todos los elementos de una lista que se ha clasificado como la otra. Para cada palabra de la lista w_i (i=1,..., N) su posición en la lista viene dada por:

$$\begin{cases} \forall\, w_i \in A\,; pos(w_i) = r\ (r\ es\ el\ rango\ de\ w_i\ en\ la\ lista\ A) \\ \forall\, w_i \notin A\,; pos(w_i) = |A| + 1 \end{cases}$$

y la permutación total de todos los elementos está dada por:

$$dist(w)_{A,B} = \sum_{i=1}^{N} |pos_A(w_i) - pos_B(w_i)| \qquad (5.5)$$

En la Tabla 5.9 está mostrando el resultado de la medida de distancia. Como podemos ver la medida de la distancia entre la lista definitiva y la de términos con alta frecuencia es también menor que la distancia entre otras listas y la lista final.

Tabla 5.9: Medidas de similitud y distancia entre la lista definitiva y otras listas

Listas de palabras vacías	Índice de Jacard	Medida de distancia
Lista de las palabras con alta frecuencia en la colección	1	6
Lista de las palabras con baja ponderación de IDF	0.9	33
Lista de las palabras generadas con entropía	0.9	28

[31] Charles Edward Spearman (Londres, 10 de septiembre de 1863 - Londres, 7 de septiembre de 1945). Psicólogo inglés. Estudió en las universidades de Leipzig, Wurzburgo y Göttingen y enseñó e investigó en la Universidad de Londres (1907 - 1931). Formuló la teoría de que la inteligencia se compone de un factor general y otros específicos. Creyó en la existencia de un factor general que interviene en todas las fases de la conducta humana y atribuyó a las capacidades específicas papel determinante en cada actividad. Escribió *The Abilities of Man* (1927), *Creative Mind* (1930) y *Psychology Down the Ages* (1937).

Referente a estos dos resultados podemos concluir que, una manera sencilla de identificar la lista de palabras vacías para un corpus de documentos persa es extraer los términos de alta frecuencia en la colección.

5.5. Conclusiones

Desde la perspectiva de un sistema de la recuperación de información, las palabras vacías no son buenos discriminadores y resultan inútiles para el propósito de la recuperación. La identificación de estas palabras es un acto preliminar para la construcción de un sistema de RI. En cuanto a la lengua persa ya hay algunas listas disponibles pero son cortas o incompletas. Casi todas las palabras vacías de la lengua persa se identificaron teniendo en cuenta de sus frecuencias en una pequeña colección de documentos. O bien, las listas fueron manualmente editadas considerando semánticamente algunas palabras para agregar o eliminar de estas listas.

En este trabajo se ha descrito un método automático para crear una lista de palabras vacías para sistemas de recuperación de información textuales en persa. Este método considera la distribución de las palabras no solo en el conjunto de colección pero también en cada documento de la colección. Además, considera el valor informativo de cada palabra. Al principio, tres listas de palabras vacías se extraen, una basada en la frecuencia de los términos en la colección, otra basada en la frecuencia inversa del documento normalizada y la última considera la medida de la entropía. La lista definitiva es la agregación de las tres listas obtenidas por cada método.

Nuestro algoritmo sólo extrae las “palabras vacías ligeras”. Decimos la “palabra vacía ligera” aquella palabra que ocurre frecuentemente en el texto, no es una palabra compuesta y contiene muy pocas letras (entre 2 y 5 letras). Mediante la combinación de estas palabras, se puede construir una lista completa de palabras vacías para cualquier sistema de recuperación de información textual en persa.

La coincidencia de nuestra lista definitiva con otra lista de las palabras vacías persas, que fue preparada utilizando el juicio de los expertos lingüísticos, es muy alta.

La superposición de nuestra lista con la lista de palabras vacías inglés ampliamente utilizada y muy conocida es también bastante alta. Esta superposición es muy alta sobre las 10 primeras palabras vacías persas.

Nuestros experimentos revelan que las palabras vacías identificadas por sus altas frecuencias en la colección tienen la mejor similitud y la menor distancia con la lista definitiva. Esto implica que una manera eficaz y sencilla de identificar las palabras vacías en una colección de documentos persas es seleccionar las palabras de uso muy frecuente en la colección.

Dado que las palabras vacías (verbales y no verbales) representan un buen porcentaje de los términos en el texto persa, reducen considerablemente el tamaño de la estructura de indexación. Por ejemplo, las 20 primeras palabras vacías pueden reducir cerca de 22% del tamaño de los términos de índice. Este resultado es aproximadamente similar al caso inglés porque las diez palabras más frecuentes en inglés generalmente representan del 20 al 30% de los *tokens* en un documento. Las 32 primeras palabras vacías ligeras tienen un gran impacto en la reducción del tamaño de índice y el conjunto de palabras vacías puede reducir el tamaño de los términos de índice alrededor de 27%.

Nuestra lista de palabras vacías fue derivada de un corpus estándar, de gran tamaño y de una grande variedad de temas. Se puede considerar como una lista general de palabras vacías persa al utilizar en el dominio del procesamiento del lenguaje natural y especialmente para los propósitos de la recuperación de información. Hay muy pocas palabras vacías que, aunque están en la lista, no pueden ser consideradas como palabras vacías en una colección general. Estas palabras son dependientes del corpus y para omitirlas, nos queda a aplicar nuestro algoritmo en diferentes corpus para generar diferentes listas finales. La agregación de estas listas nos permite identificar una lista de palabras vacías más general para la lengua persa.

Capítulo 6

Conclusiones y Trabajo Futuro

Resumen

En este último capítulo se presentan las conclusiones y aportaciones generales derivadas de la investigación desarrollada en esta tesis doctoral. Además, se enumeran algunas de las líneas de investigación que quedan abiertas y que se pretenden cubrir en trabajos futuros.

6.1. Conclusiones

El objetivo perseguido en este trabajo de tesis ha sido analizar la recuperación de información en persa e identificar los factores que pueden afectar la eficiencia de los sistemas de recuperación de información con documentos en dicho idioma. El trabajo desarrollado con esta investigación ha confirmado la hipótesis planteada inicialmente, revelando como una conclusión general que existe la necesidad de representar apropiadamente en los sistemas de RI los documentos escritos en lengua persa, para lograr una eficaz y eficiente recuperación de la información que contienen. La representación de documento es el conjunto de operaciones que se hacen sobre el contenido del mismo y su objetivo es convertir el contenido del documento a los términos de índice más consistentes para el proceso de la búsqueda.

Se considera conveniente interpretar el problema de la representación de documentos en persa en un sistema de RI tomando en cuenta dos aspectos fundamentales; la naturaleza de este idioma y el número insuficiente de investigaciones realizadas en el dominio del procesamiento del texto en esa misma lengua.

A continuación, se detallan los resultados del análisis realizado y sus conclusiones.

El primer paso realizado ha sido un análisis lingüístico, es decir, la revisión de la lengua y su escritura desde el punto de vista de procesamiento del texto para un sistema de RI. El persa es una lengua que requiere de tratamiento complejo y desafiante en el campo de procesamiento del lenguaje natural debido a las características particulares de su morfología. Las diferentes formas de la escritura, las ambigüedades en el texto escrito, la dispersión en posición alfabética diferente y no estándar de la ortografía son algunos de los principales problemas a unificar. Esto podría lograrse a través de la normalización del texto al convertir todas las formas de escritura posibles en una única norma estándar antes de su procesamiento. Esta tarea se llama "pre-normalización" o "preparación del texto" y es una tarea adicional a realizarse antes del proceso de generación de los términos asociados a un documento en un sistema de RI. La pre-normalización del texto se hace mediante herramientas que consideran las características de la lengua y que, en el caso de la lengua persa, necesitan bases de datos para manejar las excepciones y la diversidad de formas de la escritura o diferentes estilos de ortografía de las palabras. La complejidad asociada al lenguaje natural cobra especial relevancia cuando necesitamos recuperar información textual que satisfaga la necesidad de información de un usuario. La variación lingüística provoca el silencio documental, es decir la omisión de documentos relevantes para una consulta dada, ya que no se han utilizado los mismos términos que aparecen en el documento.

La segmentación del texto y la definición del límite de las palabras son tareas muy complejas para la lengua persa. Las letras tienen de uno a cuatro formas de escritura distintas. Cada forma se utiliza en función de la posición de la letra dentro de la palabra que puede ser inicial, media, final y aislada. Hay varias formas de escritura del texto que difieren en el estilo de escritura de palabras usando o eliminando los espacios dentro o entre las palabras utilizando diversas formas de caracteres. De manera que la correcta tokenizatión y la conversión de estas formas y estilos en una única norma es un paso necesario previo en la construcción de los sistemas de RI con documentos escritos en persa ya que las palabras identificadas serán los términos de índice para las etapas posteriores.

La segunda parte de esta investigación fue elaborar el estado del arte de la recuperación de información en persa. Los trabajos realizados por los investigadores en

el ámbito de la RI son muy dispersos y, a veces, muy similares. La primera necesidad de materiales que ayudan los investigadores a desarrollar y probar las herramientas adecuadas en el campo de RI es tener o acceder a colecciones estándares de prueba. La elección de una colección adecuada es de suma importancia a la hora de evaluar un sistema, ya que únicamente así tendremos la convicción de que los resultados obtenidos son fiables y representativos. La mayoría de las colecciones de prueba en persa son pequeñas colecciones de documentos que fueron construidas por cada grupo de investigadores para probar las herramientas desarrolladas por ellos mismos y generalmente no están disponibles al público.

Entre todas las colecciones de prueba, el corpus Hamshahri es la única colección estándar, fue construida hace poco tiempo y se utiliza como una fuente de documentos persas fiable y útil para los investigadores de RI. Esta colección se compone de artículos de prensa, extraídos de un periódico iraní Hamshahri (ciudadano, en español) entre los años 1996 a 2003. La colección de Hamshahri fue construida según las especificaciones de TREC. Su tamaño es relativamente más largo que otras colecciones. Las dos últimas versiones (CLEF2008 y CLEF2009) tienen alrededor de 320.000 documentos y 150 temas en formato XML.

Desde la perspectiva de un sistema de la recuperación de información, las palabras vacías no son buenos discriminadores y resultan inútiles para el propósito de la recuperación. La identificación de estas palabras es un acto preliminar para la construcción de un sistema de RI. En cuanto a la lengua persa ya hay algunas listas disponibles pero son cortas o incompletas. Casi todas las palabras vacías de la lengua persa se identificaron teniendo en cuenta sus frecuencias en una pequeña colección de documentos. Algunas de estas listas fueron editadas manualmente para agregar o eliminar de ellas algunas palabras, tomando en cuenta la semántica de las mismas.

En el campo de la RI se han realizado muchos experimentos para determinar el valor de la lematización en el proceso de la recuperación. La lematización parece, en consecuencia, fuertemente dependiente del idioma en que se encuentran escritos los documentos y consultas, de manera que resulta difícil aplicar algoritmos diseñados para un idioma a información en otra lengua diferente. Para la lengua persa han sido

desarrollados algunos lematizadores. La mayoría de estos analizadores lingüísticos utilizan la estructura de las palabras y las reglas morfológicas de la lengua persa, otros usan la base de datos y no hay ninguno que utiliza el enfoque estadístico. Hay algunos algoritmos que son lematizadores ligeros y tratan sólo los afijos y no todos los signos plurales. Existen otros algoritmos que son más completos y tratan también los verbos y utilizan una base de datos o tablas internas para manejar las excepciones. Por lo general, los lematizadores existentes no tratan completamente los prefijos y los verbos.

En el idioma persa, las palabras se construyen generalmente a partir de las formas imperativas de los verbos. Por lo tanto, desde un punto de vista lingüístico, el primer paso en la extracción de la raíz es encontrar el modo imperativo de la palabra. En general, obtener el modo imperativo no resulta fácil ya que hay infinitivos irregulares. Hay también muchas excepciones, por ejemplo como los plurales irregulares de los que no podemos extraer sus singulares. Entonces, la construcción de un algoritmo de lematización basado totalmente en reglas no es fiable. Un lematizador eficiente para la lengua persa necesita algoritmos muy potentes con aplicación de bases de datos que consideran todas las excepciones de la lengua.

En cuanto a los modelos de la RI, los modelos aplicados a los documentos persas son los mismos modelos que fueron creados por los investigadores a lo largo del tiempo para otras lenguas. En algunos trabajos, los modelos fueron aplicados en diferentes sistemas de RI. Cada sistema tenía su propia colección de documentos, sus consultas y sus métodos de evaluación. Como no se han evaluado todos estos modelos en la misma condición de evaluación resulta difícil saber qué modelos se ajustan mejor a la recuperación de documentos escritos en persa. Pero de todos estos experimentos tenemos la fuerte convicción de que ambas operaciones, la eliminación de palabras vacías y la lematización del texto, ayudan a mejorar la eficacia de recuperación para cualquier modelo.

De los modelos que fueron aplicados sobre los documentos de un corpus estándar, es decir la colección Hamshahri, se tiene algunas conclusiones, destacando las siguientes:

- Considerando la precisión de los 20 primeros documentos recuperados sin tener en cuenta el valor de exhaustividad, el rendimiento del modelo difuso es mejor que el modelo de espacio vectorial aplicando a los documentos en persa.
- El peor enfoque para recuperar documentos persas es el modelo de espacio vectorial sin eliminar las palabras vacías. Eso se puede explicar que el problema de no eliminar las palabras vacías radica en que, con sistemas que atribuyen pesos a términos, y que operan con éstos, las palabras vacías introducen un factor de ruido considerable.
- Basándose siempre sobre la precisión de los 20 primeros documentos recuperados, podemos decir que el modelo de análisis del contexto local mejora marginalmente en comparación con el modelo de espacio vectorial. Esto podría ser debido al hecho de que el método de ponderación *Lnu.ltu* está funcionando bien en el idioma persa. Otra explicación también podría ser que los parámetros utilizando por el método de *LCA* son los mismos que se utilizan en TREC y puede ser que necesite una modificación de los parámetros para la colección de Hamshahri.
- Otra conclusión es que el modelo de 4-gramas basado en el espacio vectorial con esquema de ponderación *Lnu.ltu* tiene mejores resultados que 3-gramas incluso la configuración basada en los términos y que esquema de ponderación *Lnu.ltu* tienen considerablemente mejor rendimiento que *atc.atc*.
- La aplicación de métodos como N-gramas (especialmente con un pequeño valor de n) disminuyen claramente el rendimiento de la recuperación en comparación con el caso que no haya eliminación de palabras vacías y tampoco la aplicación de ningún lematizador. Eso se puede explicar por el hecho de que las palabras vacías persa tienen muy pocas letras y la aplicación de N-gramas (N de pequeño valor) o trunc-3 será generalmente sobre las palabras vacías que no son palabras significativas. Otra explicación es que la lematización permite disminuir la presencia en el texto de palabras con muy pocas letras. Estas palabras pueden ser los sufijos como por ejemplo los signos plurales que están separados de las palabras originales en la fase de tokenización del texto.

La tercera etapa del análisis ha sido una evaluación de la RI de los documentos en persa disponibles en la Web. Con el crecimiento de Internet en la sociedad iraní, la recuperación de documentos relevantes en persa es un gran desafío para los usuarios de Internet. Para conocer los motivos que afectan al rendimiento de los buscadores web en la recuperación de información relevante se ha evaluado el rendimiento y la calidad del buscador Google que se utiliza casi por el 92% de los usuarios iraníes. Para lograr nuestros experimentos, hemos construido un sitio web con documentos de la colección Hamshahri. Para buscar las páginas web (documentos relevantes) respecto a una consulta, utilizamos la versión avanzada de Google en línea. Le confiamos al buscador Google cien consultas (los temas del corpus) y guardamos las páginas web recuperadas. Después se comparó los resultados obtenidos con los documentos relevantes del corpus. La determinación del rendimiento de la recuperación se basó en la medida de la precisión y exhaustividad de los documentos recuperados por el buscador. Estas medidas fueron comparadas con los valores de referencia y los resultados obtenidos nos indican que los documentos recuperados no son muy relevantes. Las principales razones son las siguientes:

- La tokenización y la segmentación de palabras por el sistema de Google están basadas en el espacio en blanco entre las palabras. No obstante, la tokenización en el texto persa es mucho más compleja. En este idioma el espacio no es un delimitador determinista y por lo general, no sirve para distinguir a una palabra de otra. En un texto escrito en persa se puede encontrar espacio en blanco dentro de una palabra, entre palabras o puede estar ausente entre algunas palabras secuenciales. Por lo tanto, hay muchas palabras que se pueden escribir con el espacio, el espacio corto o sin espacio.
- En Google, a partir de las palabras claves del contenido del sitio web, se deduce que las palabras vacías en persa no son eliminadas en el análisis del texto y se consideran como términos de índice por el buscador. Las palabras vacías tienen un valor muy bajo de discriminación porque conducen a documentos irrelevantes cuando se consideran para el propósito de la búsqueda. La exclusión de estas palabras se reduce el tamaño del índice y generalmente mejora la eficacia de recuperación.

- Las variantes encontradas de palabras claves del contenido del sitio web no son correctas. La constitución de las variantes de una palabra se hace normalmente durante la operación de lematización del texto. Se considera que, en particular, los tratamientos específicos para el idioma persa y su análisis morfológico no son tomados en cuenta por el buscador. Los algoritmos de Google al nivel de las reglas gramaticales de la lengua persa deben ser mejorados para poder construir correctamente las variantes de palabras del texto.

De acuerdo con lo precedente, se deduce que el sistema de Google representa un documento persa utilizando su contenido completo. Es lo que se denomina representación a texto completo del documento. Efectivamente es la forma más completa de representar un documento, pero implica un coste computacional muy alto para colecciones grandes de documentos. En este caso, cada documento se puede representar por todas sus palabras, tanto nombres, como verbos, adjetivos, adverbios, etc. A pesar de ello, no todos los términos poseen la misma utilidad para describir el contenido de un documento. De hecho, hay términos más importantes que otros, pero no es tarea fácil decidir la importancia de cada término.

La conclusión general de esta evaluación es que el buscador Google tiene problemas de relevancia al nivel de la extracción de información en persa. Estos problemas se deben a las carencias en el análisis morfológico del texto persa. El análisis morfológico de una palabra es identificar sus morfemas, sus variantes, su modelo y su raíz. Se considera que Google debe mejorar las operaciones que corresponden a la representación de documentos persas teniendo en cuenta de la estructura y gramática de la lengua persa.

Un caso donde se tienen mejores valores de precisión de los documentos recuperados por Google es cuando hay nombres propios en las consultas solicitadas. Es decir que, en este caso, la similitud entre las consultas y los documentos es un factor importante para recuperar documentos relevantes y la presencia de nombres propios lo facilita.

La última parte de este trabajo es una aportación que ayuda a eliminar un pequeño obstáculo en el proceso de la representación de documentos en persa en los sistemas de RI. Así que se ha desarrollado un método automático para crear la lista de palabras vacías persa para sistemas de RI textuales. Este método considera la distribución de las palabras no sólo en la colección sino también en cada documento de la misma; además, considera el valor informativo de cada palabra. Al principio, se extraen tres listas de palabras vacías, una basada en la frecuencia de los términos en la colección, otra basada en la frecuencia inversa del documento normalizada y la última considerando la medida de la entropía. La lista definitiva es la agregación de las tres listas obtenidas por cada método.

El algoritmo que se propone es esta trabajo sólo extrae las palabras vacías ligeras. Se denomina "palabra vacía ligera" aquella palabra que ocurre frecuentemente en el texto, no es una palabra compuesta y contiene muy pocas letras (entre 2 y 5 letras). Mediante la combinación de estas palabras, se puede construir una lista completa de palabras vacías para cualquier sistema de recuperación de información textual en persa.

La coincidencia de esta lista definitiva es muy alta con otra lista de palabras vacías persas que fue preparada utilizando el juicio de los expertos lingüísticos. La superposición de la lista que aquí se propone es también bastante alta con una lista de palabras vacías en inglés ampliamente utilizada y muy conocida. Esta superposición es muy alta sobre todo en las 10 primeras palabras vacías persas.

Los experimentos realizados en esta investigación revelaron que las palabras vacías identificadas por sus altas frecuencias en la colección tienen mayor similitud y menor distancia con la lista definitiva. Esto implica que una manera eficaz y sencilla para identificar las palabras vacías en una colección de documentos persas es seleccionar las palabras de uso muy frecuente en la colección.

Dado que las palabras vacías (verbales y no verbales) representan un buen porcentaje de los términos en el texto persa, reducen considerablemente el tamaño de la estructura de indexación. Por ejemplo, las 20 primeras palabras vacías pueden reducir

cerca de 22% del tamaño de los términos de índice. Este resultado es aproximadamente similar al caso inglés porque las diez palabras más frecuentes en inglés generalmente representan del 20 al 30 por ciento de los *tokens* en un documento. Las 32 primeras palabras vacías ligeras tienen un gran impacto en la reducción del tamaño de índice y el conjunto de palabras vacías puede reducir el tamaño de los términos de índice alrededor de 27%.

6.2. Trabajo Futuro

Los resultados del trabajo presentado en esta memoria marcan un punto de partida para el desarrollo de nuevas aproximaciones en el campo de RI aunque pueden ser también útiles en el campo de procesamiento del lenguaje natural y su aplicación al procesamiento automático de la información en la lengua. Sin embargo, es preciso seguir profundizando estas investigaciones sobre el idioma persa para reducir en lo posible las diferencias observadas con relación a los sistemas de recuperación de información en otros idiomas.

Debido a las particularidades de la lengua persa, así como a sus diferencias morfológicas y lingüísticas en comparación con otros idiomas, como el inglés, el diseño de un sistema de recuperación de información en persa requiere consideraciones especiales. Hay gran cantidad de documentos en persa disponibles en formato digital y mucho más se crean todos los días. Por lo tanto, hay una necesidad de implementar sistemas de recuperación de información con alta precisión para este idioma.

A continuación, se proporciona algunas sugerencias sobre los problemas que deben ser resueltos o abordados en el futuro para construir sistemas de RI potentes que permitan recuperar documentos más relevantes a partir de consultas realizadas en lengua persa.

- Creación de un corpus más grande, y general que sea la base para el resto de avances de la RI en persa.

- Pre-normalización y estandarización del texto: Desarrollar herramientas que permitan convertir todas las formas de escritura de una palabra en una sola norma antes de procesar el texto.
- Tokenización: Construcción de nuevos tokenizadores o mejorar los ya existentes, de manera que consideren todos las dificultades que se presentan en la recuperación de información en lengua persa al momento de distinguir palabras en un texto.
- Las palabras vacías: La lista de palabras vacías propuesta en este trabajo fue derivada de un corpus estándar, de gran tamaño y de una gran variedad de temas. Sin embargo, hay pocas palabras vacías que, aunque están en la lista, no pueden ser consideradas como palabras vacías en una colección general. Estas palabras son dependientes del corpus y para omitirlas, se recomienda aplicar el algoritmo que aquí se propone a diferentes corpus para generar diferentes listas. La agregación de esas listas permitirá construir una lista de palabras vacías más general.
- Lematización: La consideración y aplicación de todas las reglas gramaticales y excepciones de la lengua son imprescindibles en el diseño de nuevos algoritmos de lematización. La implementación de estos lematizadores deben tratar los prefijos, todos los tiempos de un verbo y posibles modificaciones de las listas de sufijos para los lematizadores ya disponibles.
- Modelos de recuperación: El modelo de lenguaje es un modelo que depende más de las características del lenguaje donde se encuentren los documentos y las consultas. Sería entonces interesante realizar estudios para mejorar sistemas basados en modelos del lenguaje para la lengua persa y comprobar estos modelos con un corpus más grande y estándar para llegar a conclusiones significativas.

Si se pudieran lograr resultados interesantes en estas líneas de trabajo entonces, se podrán construir sistemas de RI más fiables, particularmente en la Web, para los usuarios que utilizan la lengua persa.

Bibliografía

Akasereh, M., & Savoy, J. (2012). Retrieval effectiveness study with Farsi. *COnférence en Recherche d'Information et Applications (CORIA)*, (pp. 25-40). Bordeaux.

Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Toward an ARABIC Stop-Words List Generation. *International Journal of Computer Applications , 46* (8).

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Journal of Knowledge-Based Systems , 22* (5), 382-387.

Amati, G., & Van Rijsbergen, C. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems , 20* (4), 357-389.

Amtrup, J. W., Mansouri Rad, H., Megerdoomian, K., & Zajac, R. (2000). *Persian-English Machine Translation: An Overview of the Shiraz Project.* New Mexico State University. Las Cruces, New Mexico: Computing Research Laboratory. Memoranda in Computer and Cognitive Science MCCS-00-319.

Ashouri, D. (1996). *باز اندیشی زبان فارسی (Repensar de la lengua persa)* . Tehran: Nashre Markaz.

Baeza-Yates, R. (2004). Challenges in the Interaction of Information Retrieval and Natural Language. *In Proc. 5 th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 445-456). Seoul : Springer Berlin Heidelberg.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* New York, USA: Addison-Wesley .

Bateni, M. (2003). *توصیف ساختاری دستوری زبان فارسی بر بنیاد یک نظریه عمومی زبان (Estructura gramatical Pérsico basado en una teoría general del lenguaje).* Tehran: Amir Kabir.

Belkin, N. J., & Croft, W. B. (1987). Retrieval techniques. *Annual Review of Information Science and Technology , 22*, p. 109−145.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems , 30* (1-7), 107-117.

Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic Clustering of the Web. *Computer Networks and ISDN Systems , 29* (8-13), 1157-1166.

Buckley, C., & Voorhees, E. M. (2000). Evaluation Measure Stability. *In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in Information Retrieval* (págs. 33–40). New York: ACM.

Cacheda Seijo, F., Fernandez Luna, J. M., & Huete Guadix, J. F. (2011). *Recuperación de Información. Un Enfoque Práctivo y Multidisciplinar.* Madrid: RA-MA EDITORIAL.

Castillo Sequera, J. L. (2010). *Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información (Tesis doctoral).* Universidad de Alcalá, Departamento de ciencias de la computaciónl.

Cavnar, W. (1995). Using An N-Gram-Based Document Representation with A Vector Processing Retrieval Model. *In Proceedings of the Third Text REtrieval Conference (TREC-3)* (págs. 269-277). Gaithersburg, Maryland: National Institute of Standards and Technology (special publication 500-225).

Chen, S. F. (1996). *Building Probabilistic Models for Natural Language (Ph.D. thesis).* Cambridge, Massachusetts: The Division of Applied Sciences; Harvard University.

Cleverdon, C., Mills, J., & Keen, M. (1966). *Factors Determining the Performance of Indexing Systems.* Cranfield: Staff publications - Cranfield Library.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edition.* New York: John Wiley.

Croft, B., Metzler, D., & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice.* Addison-Wesley Publishing Company.

Davarpanah, R. M., Sanji, M., & Aramideh, M. (2009). Farsi lexical analysis and stop word list. *Library Hi Tech , 27* (3), pp. 435 – 449.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science , 41*, 391-407.

Di Nunzio, G. M., & Ferro, N. (2008). Results of the PERSIAN@CLEF Track in CLEF 2008: Ad Hoc Track Overview. *CLEF 2008 Workshop.* Aarhus, Denmark: Springer Berlin Heidelberg.

El-Khair, A. (2006). Effect of Stop Words Elimination for Arabic Information Retrieval: A comparative Study. *International journal of Computing & Information Sciences , 4* (3), 119-133.

Esmaili, K. S., Abolhassani, H., Neshati, M., Behrangi, E., Rostami, A., & Nasiri, M. M. (2007). A Test Collection for Evaluation of Farsi Information Retrieval Systems. *International Conference on Computer Systems and Applications, 2007. AICCSA '07. IEEE/ACS* (pp. 639 - 644). Amman: IEEE.

Estahbanati, S., Javidan, R., & Nikkhah, M. (2011). A New Multi-Phase Algorithm for Stemming in Farsi Language Based on Morphology. *International Journal of Computer Theory and Engineering , 3* (5), 623-627.

Ferro, N., & Peters, C. (2009). CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. *Workshop on Cross-Language Information Retrieval and Evaluation* (pp. 13-35). Corfu, Greece: Springer Berlin Heidelberg.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval Data Structure and Algorithm.* New Jersey, USA: Prentice Hall.

Garamaleki, F. M., & Oroumchian, F. (2002). An evaluation of retrieval performance using farsi text. *in Proc. of the First Eurasia Conference on Advances in Information and Communication Technology*, (pp. 29–31). Tehran - Iran.

García Gutiérrez, A. L. (1985). Normalización general y documental: concepto, historia e instituciones. (Documentación de las ciencias de la información IX, Universidad Complutense de Madrid,). Madrid, España.

Garg, D., & Sharma, D. (2012). Information Retrieval on the Web and its Evaluation. *International Journal of Computer Applications , 40* (3), 26-31.

Hajjar, A., Hajjar, M., Lebbos, G., Zreik, K., & El-Sayed, M. (2014). Performances of the Most Popular Search Engines in Arabic Language. *International Journal of Computer Theory & Engineering , 6* (1), 4.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval (Ph.D. Thesis).* Enschede, The Netherlands: Centre for Telematics and Information Technology.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science , 47* (1), 70-84.

Internet World Stats. (2013). *Internet Users - Top 20 Countries - Internet Usage*. Recuperado el 12 de Mayo de 2015, de Internet World Stats: http://www.internetworldstats.com/top20.htm

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (págs. 341-348). Stroudsburg, PA, USA: Association for Computational Linguistics .

Jadidinejad, A., Mahmoudi1, F., & Dehdari, J. (2009). Evaluation of perstem: a simple and efficient stemming algorithm for Persian. *Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation: text retrieval experiments* (pp. 98-101). Corfu - Greece: Springer Berlin Heidelberg.

Kashefi, O., Mohseni, N., & Minaei, B. (2010). Optimizing Document Similarity Detection in Persian Information Retrieval. *Journal of Convergence Information Technology , 5* (2), 101-106.

Korfhage, R. R. (1997). *Information Storage and Retrieval.* New York: John Wiley & Sons.

Kowalski, G. (1997). *Information Retrieval Systems: Theory and Implementation.* Norwell, Massachusetts: Kluwer Academic Publishers.

Kraaij, W., & Pohlmann, R. (1995). Evaluation of A Dutch Stemming Algorithm. *The New Review of Document and Text Management , 1*, 25-43.

Kucera, H., Francis, W. N., & Mackie, A. W. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar.* New York: Houghton Mifflin.

Lovins, J. B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics , 11*, 22-31.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development , 1* (4), 309-317 .

Mahootian, S. (1997). *Persian.* London: Routledge.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Mayfield, J., & McNamee, P. (2003). Single n-gram stemming. *Proceedings of the 26th annual international ACM SIGIR onference on Research and development in informaion retrieval* (pp. 415-416). New York: ACM.

Mazdak, N. (2004). *A Persian text summarizer (Master Thesis).* Stockholm: Department of Linguistics - Stockholm University.

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval , 7* (1-2), 73-97.

Megerdoomian, K. (2004). A Semantic Template for Light Verb Constructions. *In Proceedings of the First Workshop on Persian Language and Computers.* Tehran.

Megerdoomian, K. (2008). *Analysis of Farsi Weblogs,Technical Report (MTR080206).* Washington DC: The MITRE Corporation.

Megerdoomian, K., & Zajac, R. (2000). *Tokenization in the Shiraz Project.* Las Cruces, New Mexico: Computing Research Laboratory (Memoranda in Computer and Cognitive Science) - New Mexico State University.

Metzler, D., & Croft, W. (2004). Combining the Language Model and Inference Network Approaches to Retrieval. *Information Processing & Management , 40* (5), 735–750.

Miller, D. R., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 214–221). New York: ACM.

Mortezai, L. (2006). مسائل زبان و خط فارسی در ذخیره سازی و بازیابی اطلاعات (Problemas de la escritura y lengua persa en guardar y recuperar información). *فصلنامه اطلاع رسانی (Iranian Information & Documentation Center*), *17* (1-2).

Myerson, R. (2013). Fundamentals of social choice theory. *Quarterly Journal of Political Science , 8* (3), 305-337.

Nayyeri, A., & Oroumchian, F. (2006). FuFaIR: a Fuzzy Farsi Information Retrieval System. *Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications,* (pp. 1126-1130). Dubai/Sharjah, UAE.

Oroumchian, F., Aleahmad, A., Hakimian, P., & Mahdikhani, F. (2007). N-Gram and Local Context Analysis for Persian Text Retrieval. *International Symposium on Signal Processing and Its Applications (ISSPA 2007)* (pp. 1-4). Sharjah - UAE: IEEE.

Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., & Raja, F. (2006). *Creating a Feasible Corpus for Persian POS Tagging.* Dubai: Technical report, no.TR3/06- University of Wollongong in Dubai.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program , 14* (3), 130-137.

Rahimtoroghi, E., Faili, H., & Shakery, A. (2010). A structural rule-based stemmer for Persian. *5th International Symposium on Telecommunications* (pp. 574 - 578). Tehran: IEEE.

Ribeiro, F. C. (Octubre de 2004). El pensamiento de Hayek y la teoría de información. *Revista Libertas (Instituto Universitario ESEADE) , 41*.

Robertson, S. E., & Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science , 27* (3), 129-146.

Robertson, S. E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing and Management: an International Journal - The sixth text REtrieval conference (TREC-6) , 36* (1), 95 - 108 .

Robertson, S., & Sparck-Jones, K. (1976). Relevance weighting of search terms. *27, No. 3*, 129-146.

Rodrigo Yuste, A. (2010). *Evaluación de sistemas de búsqueda y validación de respuestas (Tesis doctoral).* Madrid: Universidad Nacional de Educación a Distancia, Escuela Técnica Superior de Ingeniería Informática .

Sadeghi, M., & Vegas, J. (2014). Automatic identification of light stop words for Persian information retrieval systems. *Journal of Information Science , 40* (4), 476-487.

Safavi, K. (1981). *درآمدی بر زبان شناسی Introducción a la lingüística.* Tehran, Iran: Bongah Tarjome va Nashr.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management , 24* (5), 513-523.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Shamsfard, M. (2011). Challenges and Open Problems in Persian Text processing. *5th Language Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, (pp. 65–69). Poznan-Poland.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). SteP-1: A set of fundamental tools for Persian text processing. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Languages Resources Association (ELRA).

Shannon, C. E. (1948). Mathematical theory of communication. *The Bell System Technical Journal , 27* (3), 379-423.

Singhal, A. (2002). AT & T at TREC-6. *25th ACM Conference on Research and Development in Information Retrieval* (pp. 35-41). ACM.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 21-29). New York: ACM.

Strzalkowski, T., & Perez Carballo, J. (1994). Recent developments in natural language text retrieval. *Proceedings of the Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215* (pp. 123-136). Gaithersburg, MD, USA: National Institute of Standards and Technology.

Sullivan, D. (2012). *Google: 100 Billion Searches Per Month, Search To Integrate Gmail, Launching Enhanced Search App For iOS.* Recuperado el 11 de Marzo de 2015, de Search Engine Land: http://searchengineland.com/google-search-press-129925

Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science , 39* (2), 92-98.

Taghva, K., Beckley, R., & Sadeh, M. (2003b). *A list of farsi stopwords.* Las Vegas: Technical Report 2003-01, Information Science Research Institute, University of Nevada.

Taghva, K., Coombs, J., Pareda, R., & Nartker, T. (2004). Language Model-based Retrieval for Farsi. *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).*

Taghva, K., Russell, B., & Sadeh, M. (2005). A Stemming Algorithm for the Farsi Language. *Proceedings. ITCC 2005 International Conference on Information Technology: Coding and Computing* (pp. 158-162). Las Vegas: IEEE.

Taghva, K., Young, R., J., C., Beckley, R., Sadeh, M., & Pereda, R. (2003a). Farsi searching and display technologies. *In Proc. of the 2003 Symp. on Document Image Understanding Technology*, (pp. 41-46). Greenbelt,MD.

Tashakori, M., Meybodi, M. R., & Oroumchian, F. (2002). Bon, First Persian stemmer. *Proceedings of the First EurAsian Conference on Information and Communication Technology* (pp. 487-494). Tehran: Springer-Verlag.

Tawileh, W., Mandl, T., & Griesbaum, J. (2010). Evaluation of five web search engines in Arabic language. *10th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 592 - 597). Cairo: IEEE.

Tsz-Wai Lo, R., He, B., & Ounis, I. (2005). Automatically Building a Stopword List for an Information Retrieval System. *Proceedings of the fifth Dutch-Belgian Information Retrieval Workshop.* De Uithof, Utrecht University, Utrecht, the Netherlands: The Journal on the Digital Information management.

Vallez, M., & Pedraza-Jimenez, R. (2007). Natural Language Processing in Textual Information Retrieval and Related Topics. *Hipertext.net , 5*.

Van-Rijsbergen, C. (1979). *Information Retrieval.* London, England: Butterworth-Heinemann.

Vilares Ferro, J. (2005). *Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en Español (Tesis doctoral).* Universidad da Coruña, Departamento de Computación.

W3Techs. (2015). *Usage Statistics of Content Languages for Websites*. Recuperado el 22 de Mayo de 2015, de Web Technology Surveys: http://w3techs.com/technologies/overview/content_language/all

Wikipedia. (2015). *Communications in Iran - Wikipedia, the free encyclopedia.* Recuperado el 4 de Abril de 2015, de http://en.wikipedia.org/wiki/Communications_in_Iran

Wikipedia. (2014). *Idioma persa - Wikipedia, la enciclopedia libre*. Recuperado el 17 de Mayo de 2015, de http://es.wikipedia.org/wiki/Idioma_persa

Wikipedia. (2013). *Moteur de recherche — Wikipédia*. Recuperado el 14 de Septiembre de 2014, de http://fr.wikipedia. org/ wiki/ Moteur_de_recherche

Wikipedia. (2012). *Persian Language-Wikipedia, the free encyclopedia*. Recuperado el 11 de Febrero de 2015, de http://en.wikipedia.org/wiki/Persian_language

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: compressing and indexing documents and images.* San Francisco, USA: Morgan Kaufmann Publishers Inc.

WordNet. (2007). *English Stop Word List in WordNet.* Recuperado el 23 de Mayo de 2015, de University of Minnesota Duluth: http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt

World Wide Web Size. (2015). *WorldWideWebSize.com, Daily estimated size of World Wide Web*. Recuperado el Marzo de 2015, de http://www.worldwidewebsize.com/

World-english. (2012). *The most common words in English*. Recuperado el 15 de Febrero de 2014, de http://www.world-english.org/english500.htm

XPO6. (2013). *List of English Stop Words - XPO6.* Recuperado el 24 de Mayo de 2015, de http://norm.al/2009/04/14/list-of-english-stop-words/

Xu, J., & Croft, W. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (págs. 4-11). New York: ACM.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control , 8* (3), 338–353.

Zamanifar, A., Minaei-Bidgoli, B., & Kashefi, O. (2008). A New Technique for Detecting Similar Documents based on Term Co-occurrence and Conceptual Property of the Text. *In Proceedings of the 3th International Conference on Digital Information Management* (pp. 526 - 531). London: IEEE.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 334-342). ACM.

Zipf, H. (1949). *Human Behaviours and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.

Zou, F., Wang, F. L., Deng, X., Han, S., & Wang, L. S. (2006). Automatic Construction of Chinese Stop Word List. *Proceedings of the 5th WSEAS international conference on Applied computer science*, (pp. 1010-1015). Hangzhou, China.

Parte III

Apéndices

Apéndices

Apéndice A: Difusión de Resultados

El trabajo de investigación desarrollado durante la realización de la presente tesis doctoral ha dado lugar a un artículo de revista ya publicado y tres otros artículos que tenemos la intención de publicar en el futuro. A continuación se detallan dichos trabajos.

El primer artículo publicado fue deducido de los resultados del Capítulo 5: Mohammad Sadeghi, Jesús Vegas, "***Automatic identification of light stop words for Persian information retrieval systems", Journal of Information Science Agosto 2014 vol. 40 no. 4 476-487,*** DOI: 10.1177/0165551514530655. La versión en línea de este artículo se puede encontrar en:

http://jis.sagepub.com/content/early/2014/04/11/0165551514530655

El segundo artículo es el resumen del Capítulo 3 de este documento. El título del artículo puede ser "***A survey of persian information retrieval***", cuya publicación está prevista en la revista "*Information Processing & Management*[32]".

El tercer artículo puede ser también un análisis completo sobra todos los algoritmos de lematización ya disponibles en la lengua persa. A demás la mayoría de los lematizadores persas que veremos en el Capítulo 3 no tratan adecuadamente la lematización de los verbos en persa. Entonces, sería interesante hacer un panorama general de los lematizadores persas y una propuesta de un algoritmo de lematización para los verbos. El título del artículo puede ser "***Overview of Persian stemmers and a stemming algorithm for verbs***". La publicación está prevista en la revista *Journal of Information Science*[33].

---

[32] http://www.journals.elsevier.com/information-processing-and-management/

[33] http://jis.sagepub.com/

El cuarto artículo será el resumen del Capítulo 4 de la tesis. El título del artículo puede ser "***How well does Google work with the persian documents?***". La publicación está prevista también en la revista "*Journal of Information Science"* para este año.

Apéndice B: Códigos de Fuente Utilizados

B.1) El programa siguiente permite separar cada documento de los archivos XML en el corpus Hamshahri. Todos los archivos XML (1.927 archivos en total) están en una carpeta. Cada archivo se procesa en su turno para estar dividido en documentos individuales. Un documento está posicionado entre las etiquetas <DOC> y </DOC> en el archivo inicial. Al final, hay 166.774 documentos.

```
import  java.io.*;
import  javax.xml.transform.*;
import  javax.xml.transform.stream.*;
import  javax.xml.xpath.* ;
import  javax.xml.parsers.DocumentBuilder.*;
import  javax.xml.parsers.DocumentBuilderFactory;
import  javax.xml.transform.dom.DOMSource ;
import  org.w3c.dom.Document;
import org.w3c.dom.NodeList;
import org.w3c.dom.Node;
import java.io.File;
import org.xml.sax.* ;
import javax.xml.parsers.DocumentBuilder;
public class XmlSplitFolder {
   // Split all not well formed xml files in a folder into mutiple xml files
      public static void main(String [] args) throws Exception {
      File folder = new File("c:/mywork/HamshahriCorpus/2003/");
      File[] listOfFiles = folder.listFiles();
```

```
        System.out.println(listOfFiles.length);
        int fileNumber = 1;
        for (int j = 0; j < listOfFiles.length; j++) {
           File input = listOfFiles[j];
           System.out.println(input);
           DocumentBuilderFactory dbf = DocumentBuilderFactory.newInstance();
           DocumentBuilder dBuilder = dbf.newDocumentBuilder();
               dBuilder.setEntityResolver(new EntityResolver()
           {
              public InputSource resolveEntity(String publicId, String systemId){
                 return new InputSource(new ByteArrayInputStream("<?xml version='1.0'
encoding='UTF-8'?>".getBytes()));
              }
           });
           Document doc = dBuilder.parse(input);
           XPath xpath = XPathFactory.newInstance().newXPath();
           NodeList nodes = (NodeList) xpath.evaluate("//HAMSHAHRI/DOC", doc,
XPathConstants.NODESET);
           int itemsPerFile = 1;
           Document currentDoc = dbf.newDocumentBuilder().newDocument();
           Node rootNode = currentDoc.createElement("HAMSHAHRI");
           File currentFile = new
File("c:/mywork/HamshahriCorpus/xml2003/"+fileNumber+".xml");
              for (int i=1; i <=nodes.getLength() ; i++) {
                 Node imported = currentDoc.importNode(nodes.item(i-1), true);
                 rootNode.appendChild(imported);
                 if (i % itemsPerFile == 0) {
                    writeToFile(rootNode, currentFile);
                    rootNode = currentDoc.createElement("HAMSHAHRI");
                    currentFile = new
File("c:/mywork/HamshahriCorpus/xml2003/"+(++fileNumber)+".xml");
                    System.out.println(fileNumber);
```

```
                }
            }
        }
    }
    private static void writeToFile(Node node, File file) throws Exception {
        Transformer transformer = TransformerFactory.newInstance().newTransformer();
        transformer.transform(new DOMSource(node), new StreamResult(new
FileWriter(file)));
    }
}
```

B.2) El programa siguiente permite de cambiar el nombre del nuevo archivo XML con el nombre del documento representando por el valor de <DOCID>. Recordamos que el valor de la etiqueta <DOCID> es el mismo que el de <DOCNO>. Este valor es el nombre del documento que existe también en el archivo de juicio de relevancia en el corpus Hamshahri.

```
import javax.xml.parsers.DocumentBuilder.*;
import javax.xml.parsers.DocumentBuilderFactory;
import org.w3c.dom.Document;
import org.w3c.dom.NodeList;
import org.w3c.dom.Element;
import org.w3c.dom.Node;
import java.io.File;

public class RenameXmlFileToDOCID {
    public static void main(String [] args) throws Exception {
        File folder = new File("c:/mywork/HamshahriCorpus/xml2003/");
        File[] listOfFiles = folder.listFiles();
            for (int i = 0; i < listOfFiles.length; i++) {
```

```
                    File xmlfile = listOfFiles[i];
                DocumentBuilderFactory dbf2 =
        DocumentBuilderFactory.newInstance();
                Document doc2 = dbf2.newDocumentBuilder().parse(xmlfile);
                NodeList nList2 = doc2.getElementsByTagName("DOC");
                Node nNode2 = nList2.item(0);
                Element eElement2 = (Element) nNode2;
                    String id =
eElement2.getElementsByTagName("DOCNO").item(0).getTextContent();
                    File newFile = new File
("c:/mywork/HamshahriCorpus/xml2003/"+id+".xml");
                    xmlfile.renameTo(newFile);

        }
   }

}
```

B.3) El programa siguiente permite de convertir un archivo XML a un archivo HTML utilizando el lenguaje XSL.

```
import java.io.*;
import javax.xml.transform.*;
import javax.xml.transform.stream.*;
import java.io.File;
import org.apache.commons.io.FilenameUtils;

public class XmlToHtml {

   // Convert all xml files in a folder to Html format with the same file name

   public static void main(String[] args) throws IOException {
```

```
        File folder = new File("c:/mywork/HamshahriCorpus/xml2003/");
        File[] listOfFiles = folder.listFiles();
        for (int i = 0; i < listOfFiles.length; i++) {
            File file = listOfFiles[i];
          try
          {
            TransformerFactory tFactory = TransformerFactory.newInstance();
            Source xslDoc = new StreamSource("hamshahri.xsl");
            Source xmlDoc = new
StreamSource("c:/mywork/HamshahriCorpus/xml2003/"+file.getName());
            String fileNameWithOutExt =
FilenameUtils.removeExtension(file.getName());
            String outputFileName = "c:/mywork/HamshahriCorpus/Data2003/"+
fileNameWithOutExt+ ".html";
            OutputStream htmlFile = new FileOutputStream(outputFileName);
            Transformer transformer = tFactory.newTransformer(xslDoc);
            transformer.transform(xmlDoc, new StreamResult(htmlFile));
          }
          catch(Exception e)
          {
            e.printStackTrace();
          }
        }
    }
}
```

A.4) El programa siguiente está escrito en XSL que nos permite diseñar las páginas Web. El contenido de la pagina (el valor de <BODY>) es el mismo que el texto del documento de Hamshahri. El título del documento (el valor de <TITLE>) es el mismo que el valor de la etiqueta <DOCNO> o <DOCID>. Hay otras informaciones como la fecha del documento y etc.

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
	xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
	<xsl:template match="/">
		<xsl:for-each select="HAMSHAHRI/DOC">
		 <html dir="rtl">
			<head>
				<meta http-equiv="Content-Language" content="fa" />
				<meta name="robots" content="index, follow" />
				<meta name="description"
content="{/HAMSHAHRI/DOC/CAT}" />
				<title>
					<xsl:value-of select="DOCNO"/>
				</title>
			</head>
			<body style="font-size: 14pt; font-family: Arial">
				<p align="left">
					<b><font color="blue"> Hamshahri corpus
document </font></b><br /><br />
					DOC ID : <xsl:value-of select="DOCID" />
					<br /><br />
					Date of Document: <xsl:value-of select="DATE"
/>
					<br /><br />
				</p>
				<xsl:value-of select="TEXT"/>
			</body>
		 </html>
		</xsl:for-each>
	</xsl:template>
</xsl:stylesheet>
```

Apéndice C: El Texto en Persa Utilizado para la Verificación de la Ley de Zipf

Aplicación de la ley de Zipf en un texto persa

Artículo: Gestión del conocimiento como una estrategia de negocios

مدیریت دانش از دیدگاه یك استراتژی تجاری

چکیده

در سالهای اخیر مدیریت دانش به یك موضوع مهم و حیاتی مورد بحث در متون تجاری تبدیل شده است. جوامع علمی و تجاری هر دو بر این باورند که سازمانهای باقدرت دانش می‌توانند برتری‌های بلندمدت خود را در عرصه‌های رقابتی حفظ کنند. منابع نقد و بررسی و چشم‌اندازهای رقابتی سازمان‌ها نشان دهندة تأثیرات این دیدگاه در عرصه‌های استراتژیك سازمان‌های تجاری است. و اگر سازمانی به راحتی نتواند شکل صحیح دانش را در جایگاه مناسب آن تشخیص دهد در عرصه‌های رقابتی با مشکل مواجه خواهد شد.

مدیریت دانش با روش استاندارد در ادارة یك شرکت تجاری متناسب خواهد بود. با اطمینان می‌توان گفت که پایه و اساس مدیریت دانش در یك دوره کوتاه‌مدت به بهره‌برداری بهینه از اطلاعات قابل دسترس و منابع موجود در یك شرکت منجر خواهد شد. در دوره‌های بلندمدت نیز می‌تواند پایه جدیدی جهت پیشرفت و توسعه در منافع تجاری باشد و مهارت‌ها را برای آینده مطمئن تقویت کند. در حقیقت می‌توان گفت مدیریت دانش برای همه شرکت‌هایی که خواهان ارتقاء و پیشرفت هستند به عنوان یك نیاز استراتژیك مطرح می‌گردد. مقالة حاضر ضمن توضیح و تبیین مدیریت دانش و همچنین نقش و اهمیت آن در فعالیت‌های مهم یك شرکت تجاری، به منظور کسب سود اقتصادی بیشتر، به مباحثی از قبیل اجزا مدیریت دانش، مدیر دانش، مفاهیم، کاربرد و شیوه‌های استفاده از این مقوله در موضوع تجارت به عنوان یك استراتژی موفق و مؤثر می‌پردازد.

مقدمه

امروزه رقابت در سطح جهانی و سرعت افزایش تغییر و توسعة مداوم در صنعت و تجارت مشکلاتی قابل توجه و فزاینده در همة سازمانها محسوب می‌شوند. آنچه مسلم است به منظور کسب فواید رقابتی بلندمدت از دیدگاه اطلاع‌رسانی و مدیریت دانش تنها تکیه بر دسترسی به منابع اطلاع‌رسانی خارجی و داخلی در روند اجرای کار کافی نیست بلکه در حال حاضر بهره‌برداری مؤثر از آنچه که در عمل با آن مواجه هستیم و نه فقط آنچه که در اختیار داریم به یك نیاز و ضرورت شغلی تبدیل شده است.

مدیریت دانش بعنوان یك استراتژی تجاری

امروزه همة مدیران مکانیزم‌های متعددی را جهت بهبود کارآیی داخلی و مواجه شدن با چالشهای مؤثر و متعدد در رقابت تجاری در اختیار دارند، اما در اصل فقط دو عامل عمده و اساسی وجود دارد که مدیران را از دیگران بعنوان فردی بی‌نظیر و توانا متمایز می‌سازد: مشتریان و کارمندان. ما می‌دانیم که کیفیت کار کارمندان، چگونگی همکاری و یاری آنها و زمینه مشترکی که تصمیم‌گیری می‌کنند عوامل بهترین را از عوامل عادی و سازمان موفق را از ناموفق متمایز می‌کند، هم اکنون بیشتر سازمانها استراتژی مدیریت دانش را بعنوان پایه اساسی توان قابل رقابت و الگوی رشد پایدار را نیز بعنوان بخشی از استراتژیهای شغلی بکار می‌برند. به هر حال، مدیریت دانش بعنوان یك استراتژی شغلی تنها وقتی به نتیجه نهایی خواهد رسید که نیازهای اساسی برآورده شده باشند. این نیازها باید در ارتباط با یك مورد کاری مشخص که از نیازهای کاری زیر اخذ گردیده‌اند در نظر گرفته شود:

- ایجاد یك سازمان برای سازماندهی مجدد، تجدید سازمان، تمرکززدایی همراه با نیازهای دیگر جهت بهبود انعطاف‌پذیری سازمانی.
- یك سازمان ممتاز که در همه سطوح نیرومند است (سازمانی متمرکز، غیرمتمرکز یا سازمانی که پراکنده باشد)
- توانایی بهبود یافته به منظور عکس‌العمل نشان دادن در مقابل نیازهای روزافزون بازار، رقابت بیشتر و پاسخ به نیازهای جدید.
- سازگاری با شرایط کاری تغییر یافته که توسط نیروهای خارجی ایجاد شده است. (بعنوان مثال: بازارها، رقیبان، مسئولان و غیره)
- سازوکارهایی که مبادلة اطلاعات و دانش را در سازمان افزایش داده و آن را تسهیل می‌کند تا آسیب‌پذیری را در زمانی که کارمندان کارهایشان را ترك می‌کنند یا زمانی که تغییرات داخلی در کارمندان ایجاد می‌شود کاهش دهد.
- از بین بردن فرآیندها و مراحل زائد کار و بهبود بهره‌برداری غیر مؤثر از منابع انسانی و اطلاع‌رسانی و استعدادها در مراحل مهم فرآیندی کار.
- دوباره‌کاری و اجرای بیهودة بنیانها و اساس مورد لزوم که بهره‌ای جز اتلاف وقت و هزینه دربر ندارد.
- برانگیختن انگیزة قوی‌تر در کارمندان برای اطلاع‌رسانی و به اشتراك گذاشتن دانش و آگاهی مؤثر.
- بهره‌برداری مؤثر از سرمایه‌گذاری در تکنولوژی اطلاع‌رسانی و زیرساخت‌های اطلاع‌رسانی.
- بازیابی اطلاعات و دانش در صورت لزوم.

اجزای اصلی مدیریت دانش استراتژیك در تصویر یك نشان داده شده است:

تجربه نشان می‌دهد که اجزاء منابع انسانی، تکنولوژی اطلاع‌رسانی و اطلاع‌رسانی که در فرآیندهای استراتژیك ملحق شده، همه باید بطور فشرده در مدیریت و فرهنگ کاری سازمان ادغام شوند و تنها این ترکیب اجزا مدیریت دانش است که توانایی سازمان را برای بهره‌برداری از تمامی امکانات اطلاع‌رسانی و دانش سازمانی موجب می‌شود.

مدیریت دانش کلاً به توسعهٔ مستمر مربوط می‌شود

اکثر سازمانها طی ده سال اخیر تغییرات اساسی کرده‌اند، نه تنها سازمانهای تجاری، بلکه سازمانهای منفرد نیز مرتباً روش کاری خود را تغییر می‌دهند. رسانه‌های خبری، ما را با این حقیقت آشنا می‌سازند که مجموعه‌های ایجاد شده اغلب غیرمنتظره و بطور اعجاب‌آوری خلاق و سازنده هستند و تغییرات باسرعت زیاد اتفاق می‌افتند تا با نیازهای مالی و تجاری تطابق داشته باشند. همگرایی تکنولوژی اطلاع‌رسانی و فرآیند جهانی شدن آن سبب می‌شود تا نیازها و ضروریات نسبت به سرعت تغییر و پیشرفت‌های سازمانی احساس شود.
مدیریت دانش بعنوان یك استراتژی کاری و شغلی بطور همزمان بر روی مرزهای چندگانه عمل می‌کند و نیز ابزاری جهت پیشرفت کلی برنامهٔ یك سازمان محسوب می‌گردد و از داخل سازمان قدرت ایجاد می‌کند تا چالشهای خارجی را برطرف نماید و این امر با بهره‌برداری از منابع دانش موجود در کارمندان، منابع اطلاع‌رسانی، تکنولوژی اطلاع‌رسانی و کاربردهای آن و همچنین ارتباط آنها با خریداران و دست‌اندرکاران بازار انجام خواهد شد. مفهوم مدیریت دانش تعیین‌کنندهٔ ابزاری برای انجام و پیشرفتهای کاری است که هر دو آنها شامل اجزای کار، توزیع، محصولات و خدمات است و کلید موفقیت آن این است که توانایی اجرای فعالانه برای توسعهٔ مستمر را دربر داشته باشد. با وجود این، مدیریت دانش برای بهبود مداوم همه فرآیندهای کاری مهم تحویل و تجارت یك ابزار استراتژیك است. و این امر بر پایهٔ این حقیقت متکی است که مدیریت دانش با مدیریت اطلاع‌رسانی مترادف نیست و باید سطح بسیار بالاتری در نظام سازمانی به اجرا درآید. این موضوع در گذشته نیز در زمینهٔ خدمات مدیریت اطلاع‌رسانی صادق بوده است.

مدیریت دانش به ایجاد فرهنگ جدید می‌پردازد

آنچه را که ما از آن بعنوان فرهنگ کاری یاد می‌کنیم معمولاً همان چیزی است که بعنوان روشی با آن سروکار داریم و نیز آنچه را که ما بعنوان کارمند در آن همکاری و مشارکت داشته باشیم. همچنین فرهنگ کاری متشکل از رفتار روابط انسانی، سطوح معمول استانداردها و ارزشها و نیز همه ابتکاراتمان (ابتکارات خودآگاه یا ناخودآگاه در یك بافت سازمانی) است، که منجر به عملکرد مثبت سازمانی از نظر حرفه‌ای و اجتماعی می‌شود. دلایل اساسی همکاری مردم و تصمیماتی که برای فرهنگ کاری گرفته می‌شود برای موفقیت (آلفا) یا شکست (اُمگا)٤ است. مثالهای متعددی از جهان کار برای نشان دادن آن نیازها و این گزارشها می‌توان یافت و ما همگی مثالهایی از فجایع شغلی کارمندان ردهٔ بالا که کارشان را ترك می‌کنند سراغ داریم. این افراد یا به یك رقیب می‌پیوندند و یا از جهان دیگر اهداف بهره‌وری را ضایع می‌کنند. در نتیجه، از لحاظ فرهنگی باید به استراتژی مدیریت دانش توجه اساسی معطوف شود. و تعجب‌آور نیست، که بطور معمول داستانهای موفقیت‌آمیز از مدیریت دانش که نشان‌دهندهٔ ایجاد یك فرهنگ کاری بنیادی، جدید یا در بیشتر موارد متفاوت نسبت به آنچه که در قبل وجود داشته است شنیده باشید. در موارد نهایی، مدیریت در سطح بالا باید جایگزین تغییر فرهنگی

گردد. بنابراین مدیریت دانش پایه و اساس سازنده برای سرعت بخشیدن به ایجاد فرهنگ کاری کنونی است، در محیط کنونی که به سرعت درحال تغییر است، شاید تاکتیك تغییرات تکاملی برای بیشتر کارها بسیار کند و زمانی که نتایج و انتظارات آن را درنظر می‌گیریم بسیار نامطمئن است. بنابراین فرهنگ کاری باید براساس استراتژیها و اهداف کار ایجاد شود و پیشرفتهای مستمر، آموزش سازمانی مستمر، به اشتراك گذاشتن دانش، باز بودن فرهنگ به مقدار زیاد، همکاری مفید میان کارمندان و اشخاص خارجی و هماهنگی که در میان مرزهای سنتی (داخلی و خارجی) ایجاد می‌شود را برانگیزاند.

مدیریت دانش به اطلاعات می‌پردازد

میزان صحیح اطلاعات در زمان مناسب مدتها است که یك عامل مهم برای انجام انواع کارها بوده است. گرچه، میزان اطلاعات موجود داخلی و خارجی زیاد است. بیهوده نیست که امروزه مدیران کاری دربارهٔ غرق شدن در دریایی از اطلاعات صحبت می‌کنند. دو برابر شدن اطلاعات موجود در هر ۸ تا ۹ ماه یکبار بر ابعاد این معضل می‌افراید. سعی بر کنترل گنجینه اطلاعات از نظر تعریف غیرممکن است. حداقل برای سازمانی که کار انجام می‌دهد بسیار پرهزینه خواهد بود. بنابراین ابزار کنترل دیگری باید پیدا شود. سؤال اصلی بیشتر این است که ما چگونه می‌توانیم زندگی کردن را تجربه کنیم و شیوه‌های کاریمان را در کنار شرایط هرج و مرج اطلاعات اجرا کنیم؟ در یك چنین شرایطی، درك اینکه کدام اطلاعات برای کارمان مهم هستند، و کدام اطلاعات بطور کلی برای کار و فرآیندهای مختلف ایجادکنندهٔ ارزش از اهمیت کمتری برخوردارند برای ما مهم هستند. جنبهٔ اطلاعاتی مدیریت دانش که شامل اطلاعات داخلی و خارجی از زمان پیدایش آن است، باید از طریق مراحل استفاده و نگهداری در سرتاسر سازمان و از جهان خارج (شرکا، مشتریان، فروشندگان و دیگران) جریان پیدا کند تا زمانی که اطلاعاتی قدیمی شود و برای آنچه که ارزش دارد کاملاً از آن بهره‌برداری شود. بسیاری از متخصصان اطلاع‌رسانی با ذخیره و نظام‌مند کردن اطلاعات برای بازیابی و استفاده مجدد از آن اقدام می‌کنند. برای یك سازمان دانش تنها تأکید بر این جنبه‌ها به تنهایی کافی نیست، بلکه اطلاعات باید بر طبق نیازهای واقعی کار از نظر کیفی کنترل شود و به فرآیندهای کاری مرتبط بپیوندد. از طریق مناسبات مشترك میان اطلاعات و فرایندهای کار تجاری انسان است که سرانجام دانش و درك جدید ایجاد می‌شود. در حقیقت می‌توان گفت در فضای ایجاد شده در زمینهٔ دانش و عقاید جدید است که اساس توسعه و تغییرات بیشتر، جهت سودآوری بیشتر کار بنا گذاشته می‌شود.

مدیریت دانش به مردم می‌پردازد

فرآیندها و روشهای مدیریت دانش توانایی دارند که با مرکز اطلاعات پیوستگی واضحی داشته باشند (بعنوان ذخیرهٔ اطلاعات، متن، تصویر، صدا و غیره) و مفهوم اطلاعات نیز در حافظهٔ انسان ذخیره می‌شود و با اعمال و رفتار انسان بیان می‌شود. در این زمینه، روابط بین منابع اطلاعاتی و اشخاصی که از آنها در ارتباطات مناسب استفاده و بهره‌برداری می‌کنند، بی‌نهایت مهم هستند. ما تاکنون شاهد این بوده‌ایم که نیروی انسانی و

کارمندان مهمترین عنصر برای کارهای مختلف هستند.
توجه داشته باشید که رقبای شما دقیقاً از همان سازوکارها و ابزارهایی استفاده می‌کنند که شما از آنها استفاده می‌کنید و از همان منابع اطلاعاتی که در دسترس شما است بدین گونه که همان اطلاعات به شیوه‌ای متفاوت مورد استفاده قرار می‌گیرند و در موقعیت دیگری قرار داده می‌شوند و به گونه‌ای متفاوت تفسیر می‌شوند که ممکن است مهمترین عامل برای ایجاد یك تمایز در محیط کار شما باشد.

مدیریت دانش به یادگیری چگونگی فراگیری می‌پردازد

آیا شما تاکنون از خودتان پرسیده‌اید که چگونه از تجربه‌های دیگران استفاده کنید؟ لازم به ذکر است زمانی که این مسئله برای یادگیری به سازمان می‌رسد، آموزش افراد با آموزش سازمان در تضاد خواهد بود و واضح و روشن است که بیشتر کارمندان از طریق شرکت و درگیری در فعالیتهای کاریشان تجربه کسب می‌کنند. این نوع مشارکت و درگیری موجب تشویق و پیشرفت کارمندان خواهد بود. از طرف دیگر در مقایسه با این می‌بینیم که بین یادگیری فردی و سازمانی که بعنوان یك بدنه اصلی یادگیری است، رابطة خودکاری وجود ندارد و سازمان به خودی خود به خاطر داشتن اعضایی با سطح علمی بالاتر و با قدرت فراگیری عالی نسبت به دیگر سازمانها برتری ندارد. در حقیقت افراد اغلب با سطح علمی بالاتر در به اشتراك گذاشتن دانش با دیگران عکس‌العمل نشان می‌دهند. به هرحال، یادگیری سازمانی به این معنا است که دانش جدید، به صورتی به سازمان برگشت داده شود که فرآیندهای تجاری، بهبود یابند و نوسازی شوند. عکس‌العمل برای کسب پیشرفت‌ها باید مستمر باشد و نیز بصورتی اجرا شود که سازمان آن را بعنوان مالکیت خود بپذیرد. بعنوان مثال بایگانی یك گزارش نهایی از یك پروژه یا درج تعداد دیدگاههایی است جهت بهبود در پایگاه داده‌های بی‌معنا و اتلاف وقت برای یك سازمان آموزشی، و در یك سازمان ابزارهای اصلی و اساسی باید بیشتر مورد استفاده قرار بگیرد. بکار بردن روشهایی که ما معمولاً به آنها بهترین روش می‌گوییم از تأثیر بیشتری برخوردارند. اغلب چنین تکنیکهایی شامل طیف گسترده‌تری از مسائل مانند جریان فرآیند کار (بهبود نحوة انجام کارها)، بهره‌برداری از تجربیات گذشته، راه و روشهای پیشنهادی برای همکاری و ارتباط، استفاده وسیع از تکنولوژی اطلاعات، مراکز مهارتهای موردنیاز و بالاخره ارزیابی مستمر و فرآیند پیشرفت که تکاملی هستند، به عبارت دیگر بهترین روشها، فرآیند کار آموزشی یا یادگیری سازمانی هستند که باید کنترل و اداره شوند. بنابراین مدیریت دانش استراتژیك یکی از ابزارهای مدیریت است که بر همه عناصر و فعالیتهای مهم تجاری به منظور کسب سود اقتصادی تأکید دارد.

مدیریت دانش به ایجاد توان از طریق دانش می‌پردازد

آموزش تنها یك فرایند واکنش‌پذیر نیست که از دانش و تجارت گذشته استفاده نماید بلکه سازمانهای آموزشی واقعی تنها از طریق کارمندانی که از مهارتهای شغلی بالایی برخوردارند، و همچنین کارمندانی که در حال آماده‌سازی، پیش‌بینی، تحت تأثیر قرار دادن و شکل دادن به فرصتهای شغلی آینده هستند، با آینده مواجه می‌شوند. در حالی سازمانهای تجاری سنتی چالشهای کاری روزانه را از طریق سازمانهایی که براساس دانش و آموزش هستند حل می‌کنند. از وجه مشخصة ساختار سازمانی، سازمانهایی که براساس دانش هستند می‌توان به

مکانهای ملاقات، عرصه‌های مبادله دانش و خلاقیت و نیز ایجاد شبکه در آن سوی مرزها اشاره کرد. سازمانی که براساس چنین ساختارهایی ایجاد می‌شود توانایی بیشتری را در حل چالشهای پیچیده‌تر که در آیندة نزدیك با آن مواجه می‌شود خواهد داشت. مفهوم مدیریت دانش فعالانه به ایجاد یك پایه استراتژیك برای یك گروه یا ساختار سازمانی شبکه‌ای کمك می‌کند.

مدیریت دانش به مهارت در تغییر می‌پردازد

اجرای یك استراتژی برای مدیریت دانش دلالت دارد بر پرسش در مورد مسائلی که به فرآیندها و روشهای کنونی کار مربوط می‌شود. اما بطور فزاینده نه تنها پرسش بلکه شناسایی روشهای مؤثرتر و جدید را نیز در انجام کار دربرمی‌گیرد. با انجام این کار، همانطور که قبلاً نیز مشاهده نموده‌ایم شما عقاید مرسوم در مورد فرهنگ گروهی، رفتار مورد قبول در میان کارمندان، نظامهای ارزشی و غیره را لمس می‌کنید.

گرچه از ابتدا تأکید بر انتظاراتی که در مورد مزایایی شغلی، نحوة ارزیابی نتایج بالقوه اجرا و همچنین تصمیم‌گیری در مورد نحوة اجرای مطلوب مدیریت دانش وجود دارد از اهمیت ویژه‌ای برخوردار است. بدون تردید یك سازمان رشدیافته تقاضاهای جدید برای پیشرفت و تغییر را بسیار آسانتر از دیگر سازمانهایی که رشد کمتری دارند بررسی می‌کند. در تصویر شمارة ۲ (مقیاس شماره۱تا ۷) درك اینکه سازمان شما پس از رشد سازمانی در کدام قسمت قرار می‌گیرد مهم است.

به‌تصویر شماره ۲ مراجعه نمائید. این تصویر نشان می‌دهد که چگونه مفهوم اجزای مدیریت دانش گام به گام اما به یك درك اساسی از دورنمای کلی و وحدت میان اعمال و نتایج تحقق می‌یابد. مدیریت دانش پایه سازمان آموزشی را ایجاد می‌کند، ما سالها سرگرم درگیری، تبادل نظر و بحث دربارة سازمان آموزشی بوده‌ایم. بنابراین باید کار خود را بهبود بخشیم و با قدرت هرچه بیشتر کار کنیم و نباید گُند عمل نمائیم، این‌ها عباراتی بوده‌اند که به ما منتقل شده‌اند. بدون تردید بسیاری از این مسائل عناصر اصلی برای ایجاد سازمان آموزشی واقعی هستند. از طرف دیگر سازمان آموزشی الگو یا ساختار سازمان نیست بلکه بطور کلی شکلی جدید از به اجرا درآوردن مدیریت دانش است که برجسته‌ترین مشخصه‌های آن ارتباطات، رفتارهای متقابل و عملکردهای انسانی است.

مدیریت دانش به تکنولوژی نمی‌پردازد...

درك این حقیقت برای اجرای مدیریت دانش لازم و ضروری است، که مدیریت دانش راه‌حل اجرای یك تکنولوژی اطلاع‌رسانی نیست و تنها نشان دادن اینکه هم اکنون یك سازمان دانش بنیاد هستیم کافی نیست. پیش از این بسیاری از فروشندگان سیستم فن‌آوری اطلاعات، توانایی بازار را در دسته‌بندی تولیدات و خدمات خود می‌دیدند و من مطمئن هستم که شما نیز با این مسئله سروکار داشته‌اید. به همان صورت که یك سیستم بایگانی، یك سیستم پست الکترونیکی یا دیگر سیستمها ، بطور خودکار فایل‌های اسناد شما را سازماندهی نمی‌کنند و یا بازیابی آنها را تسهیل نمی‌کنند و کارآیی آن را نیز بهبود نمی‌بخشند یا هرآنچه را که انتظار دارید حل نمی‌کنند، این نوع سیستم مدیریت دانش نیز مشکلات اصلی شما را حل نمی‌کند.

تکنولوژی نیز اهمیت دارد

اما از طرف دیگر باید توجه داشته باشیم که مدیریت دانش فن‌آوری بسیار قوی دارد که جزء لاینفك آن است. امروزه ایجاد سازمانهای کاری مؤثر با حذف فن‌آوری اطلاعات مانند آن است که به آینده پشت کرده باشید. در واقع مدیریت دانش یك مدیریت استراتژیك است. و لازمه‌اش نیز آن است که مدیریت عالی منحصراً از فرصت‌های ارائه شده توسط فنآوری اطلاعات برای اهداف کاری بهره‌برداری کامل را بنماید. همچنین به این امر باور داشته باشید که، تمامی رقیبان شما از پیش اقدام به این کار می‌کنند و این امر برای مدیران کاری یك سازمان یك نیاز بشمار می‌رود. از سوی دیگر براین باوریم که تکنولوژی اطلاع‌رسانی تنها بعنوان ابزار حمایتی برای فرآیندها و فعالیتهای کاری یك استراتژی ضعیف است. فرصت‌های جدید ارائه شده توسط فنآوری اطلاعات نیازمند بهره‌برداری گسترده است و این امر نیز نیازمند آن است که فنآوری اطلاعات به طور کامل بعنوان جزء اصلی در فرآیندهای کاری گنجانده شود. مدیران کاری نسبت بر این مسئله آگاهی دارند که این فنآوری اطلاعات با اجرای مدیریت دانش در سازمان فواید استراتژیك به دنبال خواهد داشت.

مدیریت دانش نیازمند رهبری مدرن است

شرط لازم جهت اخذ نتایج موفقیت‌آمیز در اجرای روند مدیریت دانش آن است که مدیریت سطح بالا دارای انگیزه و پیشرو، در کار باشد و این موضوع اغلب در مدیریت که باید فرآیند را هدایت و راهنمایی کند بیان می‌شود، اما در مورد مدیریت دانش همان‌طور که قبلاً شرح دادیم در وجود پیکره سازمان نفوذ می‌کند و شالوده‌های مدیریت دانش بر همه فرآیندهای کار، رفتار، فرهنگ و ارزشها غلبه می‌کند و این عوامل مذکور کارمندان را در همهٔ سطوح و قسمت‌های متنوع سازمانی را دربرمی‌گیرد. بنابراین ضرورت و لزوم وجود مدیریت سطح بالا در یك سازمان و نظام یافته بدان معناست که ضعف‌ها و شکاف‌هایی که به طور آشکار در یك سازمان وجود دارد و نمود پیدا می‌کند بی‌درنگ باید نسبت به حل عوامل مذکور اقدام گردد و در سازمان امکان این که اعمال یا دیدگاهها بر ضد تغییر برانگیخته شوند وجود دارد. در واقع کارمند انعطاف‌پذیری که مایل به تغییر و تحول سازنده در ساختار کلی نظام مذکور است کارمندی است که انگیزه‌های او برانگیخته شود و مورد تشویق قرار گیرد در این صورت استفاده از توان بالقوه وی آشکارتر می‌گردد و این به معنای قبول حق مسئولیت‌پذیری در انجام وظایف محوله و شناخت عمقی بهینه در اجرای پروژه کار اداری است و باعث توسعهٔ بیشتر در کار می‌گردد. به عبارت دیگر تصمیمات باید در زمان معین و سطح مقتضی اتخاذ گردد. در واقع نظام‌های رسمی آنچنان که ما امروزه از آنها شناخت داریم موقعیت و قدرت را متزلزل می‌کند. بدین ترتیب نظام‌های مشوق و براساس پاداش که باید مورد استفاده قرار گیرند از اهمیت زیادی برخوردار می‌شوند. البته لزوماً این طور نیست اما واضح است که این حقایق می‌تواند منجر به ایجاد یك فلسفه مدیریت کاری کاملاً متفاوت شود و همچنین ممکن است این شیوهٔ مدیریت جدید، گونه‌ای متفاوت از کارمندان را نسبت به آنچه که ما امروزه، اغلب شاهد آن هستیم بطلبد. این مبحث مسئله‌ای در خور توجه و تفکر برای هر فردی است که درحال حاضر فرآیند انتخاب یك استراتژی مدیریت دانش است.

مدیریت دانش مناسب تمام مشاغلی است که می‌خواهند بهترین باشند

ما مکرراً با این پرسش مواجه می‌شویم که آیا حقیقتاً مدیریت دانش می‌تواند مناسب با نوع کار و شرکت باشد؟ بررسی‌های بعمل آمده نشان می‌دهد که میزان دانش و مهارت اغلب کارها طی ۲۰ سال گذشته به میزان قابل توجهی افزایش یافته است. حتی تولیدات سنتی و شرکت‌های وابسته به صنعت در اروپا نیز شاهد چنین افزایشی به میزانی بالغ بر ۳۰ تا ۷۰ درصد بوده‌اند. گروه گارتنر بیان کرده است که مدیریت دانش با روش استاندارد در اداره یك شرکت تجاری متناسب خواهد بود. با اطمینان می‌توان گفت که پایه و اساس مدیریت دانش در یك دورهٔ کوتاه‌مدت به بهره‌برداری بهتر از اطلاعات و منابع دانش در دسترس یك شرکت منجر خواهد شد. در دوره‌های بلندمدت نیز می‌تواند پایه جدیدی جهت پیشرفت و توسعه در منافع تجاری باشد و مهارتها را برای آیندهٔ مطمئن تقویت کند. در حقیقت می‌توان گفت مدیریت دانش برای همه شرکت‌هایی که خواهان ارتقاء پیشرفت هستند به عنوان یك نیاز استراتژیك مطرح می‌گردد. به هر حال مدیریت دانش در دو وجه به نظر می‌رسد و قابل تأمل است.

وجه اول: خیلی ساده است

ما قبلاً شاهد این قضیه بوده‌ایم که مدیریت دانش در موارد ذیل بعنوان یك استراتژی تجاری مطرح می‌گردد:

- با ایجاد قدرت رقابتی از طریق بهره‌برداری پیشرفته از آنچه که پیش از این در تجارت شناخته شده است.
- مهار تغییر یا بهبود مستمر و مداوم روش کار (بهترین شیوه‌ها) و کیفیت کالاها و خدمات.
- ارائه خدمات بهتر به مشتریان و بازار با استفاده از نیروی انسانی و همه منابع ساختاری در دسترس.
- آموزش مستمر در همهٔ سطوح از جمله فردی، گروهی و سازمانی.

وجه دوم: بسیار مشکل نیز هست

اگرچه مدیریت دانش بسیار ساده به نظر می‌رسد اما شرکتهایی که سعی دارند تا به شرکتهایی دانش بنیاد تبدیل شوند با مشکلات اساسی روبرو هستند. نخست اینکه پاسخ صحیحی برای همه مشکلاتی که با آن مواجه هستند وجود ندارد و همچنین علاوه براین دستورالعمل خاصی نیز جهت اجرای روش‌ها و مفاهیم مدیریت دانش موجود نمی‌باشد. بعنوان یك استراتژی برای توسعه و پیشرفت تجارت، باید در ساختارها و ارزش‌های اساسی سازمان تغییر ایجاد شود تا رقیبان متمایز شوید. لازم به ذکر است که مدیریت دانش جهت اجرا نیازمند عوامل ذیل می‌باشد.

- درك و پذیرش ارزش اطلاعات و دانش بعنوان یك ابزار استراتژیك.
- داشتن یك گروه مدیریتی که نسبت به اجرای مدیریت دانش وفادار باشند.
- داشتن قابلیت و تمایل به تغییرات
- داشتن انگیزه و علاقه جهت بهتر شدن
- تمایل به درگیر کردن کارمندان در فرآیند کار

- اعتقاد به این امر که کارمندان از توانایی خود بطور کامل بهره‌برداری نکرده‌اند
- پذیرش يك سیستم باز با توجه به سهیم شدن در اطلاعات و دانش.