Recuperación de textos electrónicos mediante índices invertidos

Recovery of electronic texts through inverted indexes

Augusto Cortez Vásquez[1]

PRESENTACIÓN: 02 DE NOVIEMBRE DE 2018 APROBACIÓN: 20 DE DICIEMBRE DE 2018

RESUMEN

Debido a la irrupción del internet y de la enorme cantidad de datos generados en la actualidad, surge la necesidad de encontrar soluciones que permitan acceder con mayor facilidad a la información almacenada en dichos datos. Cuando un usuario busca información sobre grandes colecciones de documentos de texto, debe encontrar aquellos documentos filtrados en función de su relevancia. Este proceso requiere del almacenamiento de datos de manera eficiente y del uso en técnicas de búsqueda de datos. La recuperación de la información (IR) es el área computacional que se encarga de estudiar temas relacionados con la búsqueda de información sobrealmacenada en documentos de la internet. La técnica de índices invertidos permite recuperar información relevante facilitando la recuperación de información a los usuarios.

Palabras clave: Recuperación de textos, recuperación de información, índices invertidos, indexación semántica latente

ABSTRACT

Due to the irruption of computers, the internet and the enormous amount of generated data existing nowadays, the need has arisen to find solutions that allow you to access the way and information about the stored among said data. When a user searches for information about large collections of text documents, they must find those documents filtered according to their relevance in such a way that they provide a solution to a need for information posed by a user. This process requires the storage of data efficiently and for use in data search techniques. The recovery of information (IR) is the computational area that is responsible for studying topics related to the search for information stored in documents and the internet. The inverted index technique allows retrieving relevant information facilitating the retrieval of information to users.

Keywords: Text retrieval, Information retrieval, inverted indices, latent semantic indexation.

1 Universidad Ricardo Palma. Facultad de Ingeniería. E-mail: cortez_augusto@yahoo.fr

1. INTRODUCCIÓN

Lo que la ciencia y la tecnología ha conseguido hasta ahora ha sido realmente grandioso. Solo tenemos que mirar a nuestro alrededor para atestiguar lo que el extraordinario poder de nuestra comprensión de la naturaleza y del conocimiento nos ha ayudado a conseguir. La categorización de textos consiste en etiquetar un texto o documento con una o varias categorías temáticas preestablecidas [4], [7]. El volumen de información en las diferentes áreas de conocimiento crece a un grado exponencial. De esto se deriva que su tratamiento, así como su almacenamiento, sea más complejo. A causa de ello se ha hecho necesario desarrollar nuevos instrumentos y herramientas que faciliten la realización de procesos de búsqueda de forma eficiente y efectiva, así como la administración de estos mismos recursos. El principio de ISL es que muchas palabras utilizadas en textos pueden tener significados similares. presente trabajo tiene como principal objetivo construir un modelo que permita la recuperación de información a partir de preguntas utilizando la técnica de listas invertidas.

Nuestro objetivo es desarrollar un modelo de recuperación mediante la utilización de índices invertidos. Manning y Ariza señalan, que esta es la tarea de IR más estándar. En él, un sistema tiene como objetivo proporcionar documentos de la colección que sean relevantes para una necesidad de información arbitraria del usuario, comunicados al sistema por medio de una consulta única iniciada por el usuario [3, 4].

2. MARCO TEÓRICO

2.1. Recuperación de información

La recuperación de información (IR) consiste en es encontrar material (generalmente documentos) de una naturaleza no estructurada (generalmente texto) que satisface una necesidad de información desde el interior de grandes colecciones (generalmente almacenadas en computadoras) [5], [7].

Manning y Ariza señalan que una necesidad de información es el tema sobre el cual el usuario desea saber más, y QUERY se diferencia de una consulta, que es lo que el usuario transmite a la computadora en un intento de comunicar la necesidad de información. Un documento es relevante si es que el usuario percibe que contiene información de valor con respecto a su necesidad de información personal [5], [7].

El modelo de recuperación booleana es un modelo para la información. Dicho modelo puede plantear cualquier consulta que tenga la forma de una expresión booleana de términos, es decir, en la cual los términos se combinan con los operadores AND, OR y NOT. El modelo ve cada documento solo como un conjunto de palabras.

2.2. Clasificación de textos

La clasificación de textos es un tema que forma parte de recuperación de información-RI (Information Retrieval) y que se enmarca en la disciplina del lenguaje de procesamiento natural. Tiene como propósito etiquetar, es decir, asignar etiquetas que indican a qué categoría o categorías corresponde el documento [1], [2], [9]. En el contexto del presente trabajo, entendemos *clasificar* como distinguir las características propias de un objeto y establecer las diferencias con otros objetos. En ese sentido, clasificar textos significa relacionar un texto con sus clases [10], [8].

2.3. Estrategias de clasificación

Existen dos estrategias para la clasificación de textos: la primera consiste en la incorporación de información semántica a la representación de textos. Es conveniente destacar que, en general, estos estudios están enfocados en documentos donde es factible, en la mayoría de los casos, disponer de una colección de entrenamiento para la tarea de desambiguación del sentido de las palabras (WSD, por las siglas en inglés para Word Sense Disambiguation). La segunda estrategia consiste en el uso de métodos de WSD basados en conocimiento que obtienen información desde recursos léxicos externos. Estudios realizados demuestran que, si bien este tipo de métodos suelen arrojar resultados de menor calidad que los obtenidos con métodos basados en corpus, constituyen, en muchos casos, la única alternativa realista si se desea utilizar información semántica en la representación de documentos [12].

2.3. Indexación semántica latente (ISL)

Para comprender más claramente la similitud de dos textos, se utiliza un método numérico llamado descomposición en valores singulares (SVD, por sus siglas en inglés) cuya función primordial es identificar patrones en las relaciones entre los términos contenidos en una colección de textos no estructurados. El principio de ISL es que muchas palabras utilizadas en textos pueden tener significados similares. La idea principal es emparejar conceptos en lugar de términos, o sea, un documento podría ser recuperado si comparte conceptos con otro que es relevante para la consulta dada [11], [12].

2.4. Índices invertidos

Un índice invertido es una estructura de datos que contiene los términos del vocabulario de una colección de documentos. Por cada término almacena una lista de registro con información referente al número de identificación (ID) de cada uno de los documentos que contienen al término y, de ser necesario, las posiciones en las que aparece [7], [12].

2.5. Modelo de recuperación de información booleano

El modelo de clasificación booleano permite la resolución de consultas en las que se emplean operadores como conjunción, disyunción y negación, entre otros, y permite encontrar los documentos que cumplen con los requerimientos del usuario. Este método asume que los documentos son sacos de palabras en los que no se tiene en cuenta el orden relativo entre los términos. Dado que los operadores lógicos son conmutativos, representa lo mismo la operación *W1 Y W2* que *W2 Y W1*

3. MÉTODO Y TÉCNICAS UTILIZADAS

Metodología

Para modelar el problema P de clasificación de textos se seguirá la siguiente metodología:

1) Identificar una muestra de documentos
2) Tokenizar los documentos en una tabla identificando el documento al que pertenecen
3) Eliminar token irrelevantes
4) Construir la lista de documentos por cada lexema

5) Determinar muestra de preguntas P
6) Determinar documentos que satisfacen P

4. DESARROLLO DE SOLUCIÓN

4.1. Consideremos la siguiente muestra de documentos M: T_1, T_2, T_3, T_4, T_5

T1: Las redes inalámbricas tienen ciertas características según rangos de frecuencias utilizados para transmitir señales electromagnéticas por puertos Ethernet.

T2: El medio de transmisión de las redes inalámbricas pueden ser ondas de radio, microondas terrestres o satélites y los infrarrojos.

T3 Las características de las redes inalámbricas mediante ondas de radio electromagnéticas oscilantes no requieren antenas parabólicas u otros dispositivos inalámbricos.

T4 Los dispositivos de internet que utilizan cables conectados a puertos del router de la red requieren antenas parabólicas conectadas a puertos Ethernet.

T5: Una red por cable conecta dispositivos a internet u otra red mediante cables. Las redes por cable más comunes utilizan cables conectados a puertos Ethernet del router de la red en un extremo y a un equipo u otro dispositivo en extremos opuestos del cable.

4.2. Eliminamos tokens no significativos

T1: redes inalámbricas características rangos frecuencias utilizando transmitir señales electromagnéticas puertos Ethernet

T2: medio transmisión redes inalámbricas ondas radio, microondas, terrestres satélites infrarrojos

T3: características redes inalámbricas mediante ondas radio electromagnéticas oscilantes requieren antenas parabólicas dispositivos inalámbricos

T4: dispositivos internet utilizan cables conectados puertos router red requieren antenas parabólicas conectadas puertos Ethernet

T5: red cable conecta dispositivos internet red mediante cables redes cables comunes utilizan cables conectadas puertos Ethernet router red extremo equipo dispositivo extremos opuestos cable

4.3. Tokenizamos documentos

Para la construcción de un índice invertido, es necesario determinar cuáles términos van a ser almacenados. Se realiza, en primer lugar, un proceso de tokenizacion de cada uno de los documentos. Este proceso consiste en dividir el documento en tokens, que son unidades conformadas por secuencias de caracteres que se encuentran agrupados y que tienen relevancia gramatical. Los tokens determinan, inicialmente, cuáles términos son almacenados en un índice invertido. Sin embargo, se pueden emplear diversas técnicas de eliminación o modificación de tokens con el objetivo de adecuar y/o reducir los términos indexados. Para ello, creamos un árbol de palabras irrelevantes. Cuando un token de un texto se encuentra en el árbol, se desecha; en caso contrario, se considera en la indexación.

Ord	Palabra	Texto
1	Redes	1
2	Inalámbricas	1

Ord	Palabra	Texto
1	Antenas	3
2	Antenas	4

Ord	Palabra	Cantidad	Texto
1	Antenas	2	3, 4
2	Cable	6	5, 5, 4, 5,5, 5

3	Características	1	3	Cable	5	3	Características	2	1, 3
4	Rangos	1	4	Cable	5	4	Ciertas	1	1
5	Frecuencia	1	5	Cables	4	5	Comunes	1	5
6	Utilizando	1	6	cables	5	6	Conecta	4	5, 4, 4, 5
7	Transmitir	1	7	Cables	5	7	dispositivo	4	5, 4, 5, 3
8	Señales	1	8	Cables	5	8	Dispositivos	3	4, 5, 3
9	Electromagnéticas	1	9	Características	1	9	Electromagnéticas	2	1, 3
10	Puertos	1	10	Características	3	10	equipo	1	5
11	Ethernet	1	11	Comunes	5	11	Ethernet	2	1,5
12	Medio	2	12	Conecta	5	12	extremo	2	5, 5
13	Transmisión	2	13	conectadas	4	13	frecuencia	1	1
14	Redes	2	14	Conectados	4	14	Inalámbricas	4	1, 2, 3, 3
15	Inalámbricas	2	15	Conectads	5	15	Infrarrojos	1	2
16	Ondas	2	16	dispositivo	5	16	Internet	2	4, 5
17	Radios	2	17	Dispositivos	4	17	Mediante	3	3, 5, 2
18	Microondas	2	18	dispositivos	5	18	Microondas	1	2
19	Terrestres	2	19	Dispositivos	3	19	Ondas	2	2, 3
20	Satélites	2	20	Electromagnéticas	1	20	Opuestos	1	5
21	Infrarrojos	2	21	electromagnéticas	3	21	Oscilantes	1	3
22	Características	3	22	equipo	5	22	Parabólicas	2	3, 4
23	Redes	3	23	Ethernet	1	23	Pueden	1	2
24	Inalámbricas	3	24	Ethernet	5	24	Puertos	3	1, 4, 5
25	Mediante	3	25	extremo	5	25	radio	2	3, 2
26	Ondas	3	26	Extremos	5	26	Rangos	1	1
27	Radio	3	27	frecuencia	1	27	red	8	4, 5, 5, 5, 1, 2, 3, 5
28	electromagnéticas	3	28	Inalámbricas	1	28	requieren	2	3, 4
29	Oscilantes	3	29	Inalámbricas	2	29	Router	2	4, 5
30	Requieren	3	30	Inalámbricas	3	30	Satélites	1	2
31	Antenas	3	31	inalambricos	3	31	Según	1	1
32	Parabólicas	3	32	Infrarrojos	2	32	Señales	1	1
33	Dispositivos	3	33	Internet	4	33	Terrestres	1	2
34	inalambricos	3	34	internet	5	34	tienen	1	1
35	Dispositivos	4	35	Mediante	3	35	Transmisión	2	2, 1

36	Internet	4
37	Utilizan	4
38	Cables	4
39	Conectados	4
40	Puertos	4
41	Router	4
42	Red	4
43	Requieren	4
44	Antenas	4
45	Parabólicas	4
46	Conectadas	4
47	Red	5
48	Cable	5
49	Conecta	5
50	dispositivos	5
51	Internet	5
52	Red	5
53	Mediante	5
54	Cables	5
55	Redes	5
56	Cables	5
57	Comunes	5
58	Utilizan	5
59	Cables	5
60	Conectads	5
61	Puertos	5
62	Ethernet	5
63	router	5
64	Red	5
65	Extremo	5
66	Equipo	5
67	Dispositivo	5
68	Extremos	5
69	Opuestos	5
70	Cable	5

36	mediante	5
37	Medio	2
38	Microondas	2
39	Ondas	2
40	Ondas	3
41	Opuestos	5
42	Oscilantes	3
43	Parabólicas	3
44	Parabólicas	4
45	puertos	1
46	Puertos	4
47	puertos	5
48	radio	3
49	Radios	2
50	Rangos	1
51	red	4
52	Red	5
53	Red	5
54	Red	5
1	Redes	1
2	Redes	2
3	Redes	3
4	Redes	5
5	requieren	3
6	requieren	4
7	Router	4
8	router	5
9	Satélites	2
10	Señales	1
11	Terrestres	2
12	Transmisión	2
13	Transmitir	1
14	Utilizan	4
15	utilizan	5
16	Utilizando	1

36	Utilizan	3	4, 5, 1

4.4. Determinamos muestra de preguntas

Pregunta: (redes, inalámbrica, AND)
redes: 4 ➔ 5 ➔ 5 ➔ 5 ➔ 1 ➔ 2 ➔ 3 ➔ 5
inalámbricas: 1 ➔ 2 ➔ 3 ➔ 3
(redes, inalámbricas): 1 ➔ 2 ➔ 3

Pregunta: (redes, inalámbrica, OR)
redes: 4 ➔ 5 ➔ 5 ➔ 5 ➔ 1 ➔ 2 ➔ 3 ➔ 5
inalámbricas: 1 ➔ 2 ➔ 3 ➔ 3
(redes, inalámbricas): 1 ➔ 2 ➔ 3 ➔ 4 ➔ 5

Pregunta: (puertos, Ethernet, AND)
Puertos: 4 ➔ 4 ➔ 5
Ethernet: 1 ➔ 5
(Puertos, Ethernet): 5

La operación de intersección es crucial: necesitamos intersecar de manera eficiente las listas de publicaciones para poder encontrar rápidamente los documentos que contienen ambas. (Esta operación, a veces, se denomina fusión de listas de publicaciones: este nombre ligeramente contraintuitivo refleja el uso del término algoritmo de fusión para una familia general de algoritmos que combinan múltiples listas ordenadas por avance intercalado de punteros a través de cada uno. Aquí estamos fusionando las listas con una operación lógica AND).

Entrada: D: diccionario de términos
T1, T2: términos a consultar

Salida R: lista de documentos
Precondición: T1, T2 ∈ D
Postcondición: R = L1 ∩ L2, donde L1 lista asociada a T1, L2 Lista asociada a T2

```
Consultar (D, D1, D2)
Inicio

    Localizamos D1 en D
    R1= Recuperamos Lista(D1)
    Localizamos D2 en D
    R2= Recuperamos Lista(D2)
    Resultado = intersección (D, R1, R2)
Fin
Intersección (D, T1, T2)
Inicio
    Resp = ϕ
    Mientras T1≠ Null  ^  T2≠ Null
            Si (T1.ID =  T2.ID)
                    Adicionar(T1.Id , Resp)
            Sino
```

```
                    Si (T1.ID <  T2.ID)
                        T1= T1.Sig
                    Sino
                        T2= T2.Sig
                    FinSi
                FinSi
        FinMientras
        Retornar Resp
Fin
```

El algoritmo utilizado es mergeSort. Si las longitudes de las listas de publicaciones son "x" e "y", la intersección toma operaciones en el orden O (x + y). Formalmente, la complejidad de la consulta es Q (N), donde N es el número de documentos en la colección.

La operación de intersección es crucial: necesitamos intersecar de manera eficiente las listas de publicaciones para poder encontrar rápidamente los documentos que contienen ambas. (Esta operación a veces se denomina fusión de listas de publicaciones: este nombre ligeramente contraintuitivo refleja el uso del término algoritmo de fusión para una familia general de algoritmos que combinan múltiples listas ordenadas por avance intercalado de punteros a través de cada uno. Aquí estamos fusionando las listas con una operación lógica AND)

La medida de frecuencia del término ($tf_{t,d}$) se calcula como el número de veces que aparece un término (t) en un documento (d) [5], [7].

Con el objetivo de otorgar un puntaje de relevancia, se utiliza la medida:

tf - idf, donde:

tf denota la frecuencia del término
idf denota la frecuencia inversa en los documentos

Para esto, se utiliza el modelo de espacio vectorial (VSM), en el cual cada uno de los documentos es representado mediante un vector en el que cada término del diccionario representa una dimensión. De esta forma, es posible calcular la similaridad entre un documento y una consulta con base a la suma de las medidas tf-idf de los términos de la consulta que se encuentran en el documento.

5. CONCLUSIONES

a) La indexación semántica latente permite recuperar textos atendiendo a su contenido conceptual y no solo a su contenido textual.
b) La recuperación mediante índices invertidos permite recuperar documentos que contienen información sobre tokens contenidos en los textos.
c) Con el objetivo de optimizar el proceso de indexación, eliminamos los token irrelevantes que no aportan significado. Estos se almacenan en un árbol de búsqueda binario. Durante el proceso de tokenizacion, cada token es buscado en el árbol. Si se encuentra, se desecha.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] M. Mendoza, "Categorización de texto en bases documentales a partir de modelos computacionales livianos", *Revista Signos*, versión ISSN 0718-934, vol. 44, n.º 77, Valparaíso, dic. 2011.

[2] A. Cortez, "Categorización de textos mediante máquinas de soporte vectorial", *Revista de Investigaciones de Sistema e Informática, RISI*, vol. 10, n.º 1, p. 33, 2013.
[3] J. Hernández, *Introducción a la minería de datos*. España: Pearson Prentice Hall, 2009.
[4] L. Leija, *Métodos de procesamiento avanzado e inteligencia artificial en sistemas sensores y biosensores.* México: Reverse Editores, 2009.
[5] C. Manning. *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.
[6] P. Muñoz, *Sistema para el reconocimiento fuera de línea de caracteres Manuscritos*. Universidad del Quindío: Grupo GAMA5 CEIFI, 2006.
[7] C. Ariza, *Sistema de recuperación de información utilizando índices combinados de términos y extracción de información*. Bogotá, Colombia: Universidad Nacional de Colombia, Facultad de Ingeniería, Maestría en Ingeniería de Sistemas y Computación, 2012.
[8] J. Palma, *Inteligencia artificial*. Madrid: Mc Graw Hill, 2008.
[9] M. Rosas. "Un análisis comparativo de estrategias para la categorización semántica de textos cortos", *Revista Procesamiento del Lenguaje Natural*, no. 44, pp. 11-18, 2010.
[10] S. Russell, *Inteligencia artificial: un enfoque moderno*. México: Edit. Pearson, 2003.
[11] A. Putri, "Word Level Auto-correction for Latent Semantic Analysis Based Essay Grading System", Department of Electrical Engineering, Faculty of Engineering Universitas Indonesia Depok, Indonesia. 15th Intl. Conf. QiR: Intl. Symp. Elec. and Com. Eng, 978-602-50431-1-6/17. pp. 234-236 Available: https://ieeexplore.ieee.org/document/8168488.[Accessed: Aug 22, 2018]
[12] R. Chowdhury, "An Approach to Generic Bengali Text Summarization Using Latent Semantic Analysis". Department of Electrical Engineering, Faculty of Engineering Universitas Indonesia Depok, Indonesia. 2017 International Conference on Information Technology, 978-1-5386-2924-6/17 . pp. 11-13 Available: https://ieeexplore.ieee.org/document/8168488.[Accessed: Aug 22, 2018] [Accessed: Aug 22, 2018]