EXPERIMENTO DE RECUPERACIÓN DE INFORMACIÓN USANDO LAS MEDIDAS DE SIMILITUD COSENO, JACCARD Y DICE

L. S. GARCÍA MONSALVE [1]

LUZ STELLA GARCÍA MONSALVE[1]

Ingeniera de Sistemas con énfasis en Desarrollo de Software de la Universidad Antonio Nariño. Especialista en Auditoría de Sistemas de la Universidad Antonio Nariño. Estudiante de la Maestría en Ciencias de la Información y las Comunicaciones de la Universidad Distrital "Francisco José de Caldas". Docente desde el año 1994.

RESUMEN

Con frecuencia resulta extremadamente dispendioso e incluso se puede pensar que físicamente es imposible recuperar información de otra manera que no sea automáticamente, debido al gran volumen de ésta, además, este proceso trae como consecuencia que al no ser preciso, la información relevante será ignorada por el afán de hacer el trabajo rápidamente. [1] Hoy en día es posible obtener mayores beneficios en dicho proceso de recuperación de información usando herramientas tecnológicas avanzadas diseñadas para tal fin. [2]

Tomando como referencia la colección documental de prueba ADI [12], se realizó un experimento que permitió almacenar en tablas los 82 documentos y las 35 consultas que ofrece la colección, para luego aplicar las técnicas de tokenización y stop words y calcular la frecuencia absoluta simple y la frecuencia inversa, para posteriormente hallar los resultados de los coeficientes Coseno, Jaccard y Dice, compararlos y determinar cuál de ellos tiene la mayor precisión.

PALABRAS CLAVE: Lenguaje natural, recuperación de información, métricas.

ABSTRACT

Often, results extremely difficult and you can even think that is impossible to retrieve information other than automatically, by the large volume of this, also, this process brings as consecuence that if it is not accurate, the relevant information will be ignored. [1] Today you can get more benefits in said process information retrieval, using advanced technological tools designed for this purpose. [2]

Taking as reference the documental collection of proof ADI [12], performed an experiment that allow store in tables 82 documents and 35 consultations that the collection offers, to then apply the tokenización techniques and Stop Words, and calculate the simple absolute frequency and the inverse frequency, to find later the results from coefficients Coseno, Jaccard and Dice, to compare and determine which has the highest Precision.

1. INTRODUCCIÓN

"El Procesamiento del Lenguaje Natural (PLN) es una sub-disciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional; el PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas o entre personas y máquinas por medio de lenguajes naturales"[1].

Desde los años 60 la generación y comprensión automática del lenguaje natural ha cobrado mayor importancia debido a que se ha enfocado en estudiar los problemas que resultan de la generación y procesamiento del lenguaje natural. [3] Hoy en día se usan estos aportes para procesar grandes cantidades de información, como ejemplo es posible citar: motores de búsqueda web en las herramientas de traducción automática, o en la generación automática de resúmenes.

2. METODOLOGÍA Y DESARROLLO

2.1 Recuperación de Información

En el proceso de recuperar información ha sido necesario fundamentarse de manera teórica y práctica en las mejoras que se han hecho al proceso de actualización de los motores de búsqueda, y la construcción y mantenimiento de grandes repositorios de información [4].

1 Rusell S. Inteligencia Artificial: Un enfoque moderno. Prentice Hall, 1996.

La Recuperación de la información consiste en que el usuario puede hacer una consulta con el fin de encontrar los documentos pertinentes o relevantes a su necesidad. [5] Al recuperar información se sigue un proceso de búsqueda de documentos que no están estructurados y que deben satisfacer una necesidad de información sobre grandes colecciones de documentos que generalmente están almacenados en computadoras.

2.1.1 Modelos de recuperación de Información

Constituyen una de las principales herramientas que facilitan la comparación entre una consulta determinada y una serie de textos sobre los cuales se realiza dicha consulta. [6]

Los **modelos de recuperación** se aplican a documentos de contenido únicamente textual. Al recuperar un documento se crea un índice determinado en función del contenido de dicho documento y que puede tener factores como por ejemplo la frecuencia con la cual aparece la palabra en el documento.[2]

2.1.1.1 Clasificación de los Modelos de Recuperación

A continuación se muestra una clasificación de los principales modelos de recuperación existentes en la actualidad.

2.1.1.1.1 Modelo de Recuperación Booleano

Este modelo se basa en la teoría de conjuntos del Algebra de Boole y fue adoptado por los primeros sistemas bibliográficos comerciales. Se basa en un algoritmo de decisión binaria, para determinar si un elemento esta o no contenido en el conjunto resultado.

2.1.1.1.2 Modelo de Recuperación Probabilístico

Este modelo se basa en la equiparación probabilística. Para ello dada una pregunta por el usuario se calcula la probabilidad de que esa pregunta tenga relación con el documento a recuperar.

2.1.1.1.3 Modelo de Recuperación basado en Espacio de Vectores

Este modelo se basa en espacios vectoriales, se asignan pesos a los términos índice de las preguntas y de los documentos. Estos pesos son usados para obtener similitud entre cada documento almacenado y las preguntas propuestas por el usuario.

El motor de recuperación de palabras similares se encuentra basado en el modelo de espacio vectorial o esquema TF-IDF (Term Frecuency – Inverse Document Frecuency), aplicado a tareas de clasificación y similitud de documentos. [7]

2 Martin, P.,Sergio, Modelos deRecuperación, disponible en:http://modelosderecuperacioni.iespana.es/

Los documentos se ubican en espacios vectoriales multidimensionales definido por los mismos términos. Así si cada término define una dimensión y la frecuencia de ese término define una escala lineal a lo largo de esa dimensión, las consultas y los documentos pueden ser representados por vectores en el espacio resultante.

Para el caso de este experimento se tiene en cuenta cada término de las consultas y los documentos y se pondera asignándoles el valor inverso de la frecuencia del término en los documentos de la colección (IDF – Inverse Document Frequency). Este valor se calcula usando la siguiente ecuación:

Ecuación 1

$$idf_t = \log(\frac{N}{n_t})$$

Donde N es el número de documentos en la colección y n_t es el número de documentos donde el término t aparece.

El peso de un término t en el vector del documento, se da por la ecuación:

Ecuación 2

$$W_{t,d} = tf_{t,d} X idf_{t,d}$$

Donde $f_{t,d}$ es la frecuencia absoluta del término t, en el documento actual.

2.2 Coeficientes Coseno, Jaccard y Dice

Con los elementos para la recuperación de documentos, se puede calcular la similitud entre un vector de pesos de los términos de la consulta q, y un vector de pesos de términos del documento, d, con las siguientesecuaciones:

Ecuación para hallar el coeficiente Coseno:

Ecuación 3

$$sim(q,d) = \frac{\sum w_{t,d}.w_{t,q}}{\sqrt{\sum_t w_{t,d}^2}.\sqrt{\sum_t w_{t,q}^2}}.$$

Ecuación para hallar el coeficiente Jaccard:

Ecuación 4

$$sim(q,d) = \frac{\sum (w_{t,d}.w_{t,q})}{\sum w_{t,d}^2 + \sum w_{t,q}^2 - \sum (w_{t,d}.w_{t,q})}$$

Ecuación para hallar el coeficiente Dice:

Ecuación 5

$$sim(q,d) = \frac{2\sum w_{t,d}.w_{t,q}}{\sum w_{t,d}^2 + \sum w_{t,q}^2}$$

2.3 Índices de Precisión y Recall

Para evaluar los métodos de Coseno, Jaccard y Dice, se usaron los índices de Precisión y Recall que permiten medir la calidad en la recuperación de información.

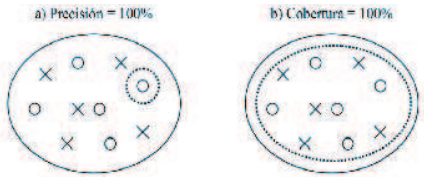

Fig.1. Precisión vs. Recall. Tomada de Enríquez, Fernando, Técnicas de Bootstrapping en el Procesamiento del Lenguaje Natural.

Los índices se calculan usando las siguientes ecuaciones:

Ecuación 6

$$Precision = \frac{documentos\ recuperados\ relevantes}{documentos\ recuperados}$$

Ecuación 7

$$Recall = \frac{documentos\ recuperados\ relevantes}{documentos\ relevantes}$$

2.4 Procedimiento

El experimento se realizó en el siguiente orden:

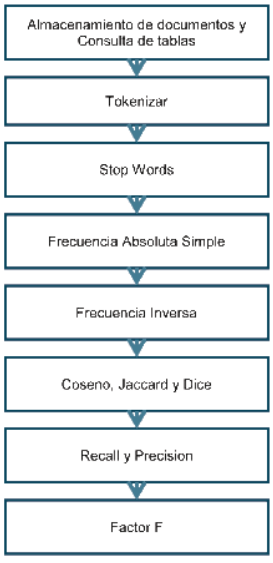

Fig. 2. Muestra el orden de los procesos que se siguieron en el experimento.

2.4.1 Almacenamiento de documentos y consultas en tablas

A través de una base de datos en PostGress[3] se generaron 2 tablas para guardar los 82 documentos y las 35 consultas de la colección documental de prueba ADI.

2.4.2 Tokenizar

Se separaron las palabras de los documentos y de las consultas. Como resultado se generaron las tablas token donde se almacenaron las palabras del resumen de los documentos y tokenconsulta donde se almacenaron las palabras de las consultas.

Para la generación de los token se utilizó la clase *StringTokenizer de java que nos permite* dividir un string en substrings o tokens, con base en otro string denominado delimitador. Para este experimento el delimitador utilizado fue el espacio en blanco " "[9].

En el resultado final se observan token innecesarios para el objetivo del experimento.

2.4.3 Stop Words

Esta técnica se utilizó en este experimento para eliminar palabras que carecen de importancia. Al analizar el contenido se puede observar que existen diversos tipos de palabras como artículos, números, símbolos, que aparecen con demasiada frecuencia y que no son relevantes para la recuperación de información, por lo tanto se puede prescindir de ellas.

3 Motor de Bases de Datos que permite manejar grandes volúmenes de datos organizados en tablas.

ID	PALABRA
1	:
2	Of
3	The
4	Within
5	1
6	.
7	A
8	For
9	And
10	2
11	It
12	Is
13	With
14	All
15	Such
24	9
25	“”
26	?
27	+
28	-
29	/
30	*

Tabla 1. Ejemplo de algunos términos usados como Stop Words y que no aportan información al experimento.

2.4.4 Frecuencia Absoluta Simple

Como parte del proceso que se debe realizar para concluir el experimento se genera la Frecuencia Absoluta Simple que muestra cuantas veces se repite una palabra en cada documento y cuantas veces se repite una palabra en cada consulta.

Token Documento 1	Frecuencia
information	2
systems	1
program	1
dissemination	1
an	1
integrated	1
system	5
manual	1
all	2
produced	1
producing	1
technical	2

Tabla 2. Ejemplo de algunos token encontrados en el documento 1.

2.4.5 Frecuencia Inversa

Una vez hallada la frecuencia simple de cada palabra (token) en cada documento y en cada consulta, se halla la frecuencia inversa del documento, que permite calcular el peso de cada palabra en cada documento y cada consulta, para ello se utiliza la ecuación 1.

Cada peso de las palabras se almacena en un vector. Para calcular los pesos se utiliza la ecuación 2.

Ejemplo

Si usamos la palabra *new* para comprobar el peso del token en los documentos. En 6 documentos se halló la palabra. Para hallar la frecuencia inversa se procede así:

$$idf = \log \frac{82}{6} = 2{,}45943$$

Documento	Token	Frecuencia
2	New	1
5	New	1
60	New	1
55	New	1
79	New	1
30	New	1

Tabla 3. Muestra los documentos en los que se halló la palabra *new* y la frecuencia absoluta simple.

El peso en el documento2 para esa palabra esta dado por:

$$W_{t,d} = 1 * 2{,}45943 = 2{,}45943$$

Debido a que los documentos generalmente contienen más palabras que las consultas y es necesario generar 2 vectores, uno para almacenar el peso de las palabras del documento1 y otro para almacenar el peso de las palabras de la consulta1 con la característica de que los vectores sean iguales. Si el vector de pesos del documento 2 contiene la palabra "new" en la posición 0 y en la misma posición del vector de pesos de la consulta no aparece esa palabra, se asigna en esa posición el peso 0; y viceversa en el caso de que esa palabra no apareciera en el documento 2. Este proceso se hace comparando el vector de pesos de la consulta 1 con cada uno de los 82 documentos, con lo cual se calcula el producto escalar de los pesos, utilizados para hallar la similitud de la consulta 1 con el documento 1.

2.4.6 Coeficientes Coseno, Jaccard y Dice

Con base en la aplicación creada en Java se obtienen los coeficientes de Coseno, Jaccard y Dice, usando la ecuación 3 para Coseno, la ecuación 4 para Jaccard y la ecuación 5 para Dice.

En la tabla 4 se muestran los resultados del coeficiente Coseno, Jaccard y Dice para la primera y segunda consultas.

Con	Doc	Coseno	Dice	Jaccard
1	4	0,187335721	0,05536413	0,02847018
1	9	0,1554845 86	0,05187763	0,02662956
1	14	0,079621333	0,01192366	0,00599759
1	16	0,079212838	0,00868468	0,00436128
1	19	0,162771235	0,04952846	0,02539307
1	22	0,230985522	0,046635	0,02387419
1	23	0,153461158	0,05410457	0,02780446
1	27	0,189074587	0,06171773	0,03184146
1	30	0,182817167	0,06209397	0,03204179
1	46	0,172254189	0,05531031	0,02844172
1	69	0,203575725	0,06511607	0,03365374
1	71	0,163694234	0,05806654	0,02990141
2	5	0,222358941	0,05705471	0,02936506
2	12	0,227501797	0,10881061	0,05753554
2	16	0,079212838	0,00868468	0,00436128
2	35	0,15640575	0,04581369	0,02344387
2	44	0,129230882	0,03149598	0,01599996
2	69	0,203575725	0,06511607	0,03365374
2	71	0,163694234	0,02975898	0,01510423

Tabla 4.Resultados de los coeficiente Coseno, Jaccard y Dice.

La siguiente tabla muestra un ejemplo de los resultados de los documentos relevantes obtenidos con los coeficientes Coseno, Jaccard y Dice, se comparan con los verdaderos relevantes.

Coseno		Jaccard		Dice		Expertos	
Con	Doc	Con	Doc	Con	Doc	Con	Doc
1	4	1	4	1	4	1	17
1	9	1	23	1	9	1	46
1	14	1	27	1	23	1	62
1	16	1	30	1	27	2	12
1	19	1	46	1	30	2	71
1	22	1	69	1	46	3	3
1	23	1	71	1	69	3	43
1	27	2	5	1	71	3	45
1	46	2	69	2	12	4	29
1	69	3	21	2	69	4	63
2	5	4	29	3	45	5	21
2	12	4	41	4	29	5	72
2	16	4	50	4	41	6	18
2	44	5	4	5	3	7	7
2	69	5	26	5	4	7	9
2	71	5	47	5	9	7	19
3	21	5	51	5	26	7	40
3	27	5	70	5	47	8	61
3	44	6	9	5	51	9	50
3	45	6	46	5	70	9	82
4	4	7	6	6	9	10	2
4	16	7	30	6	46	10	11
4	22	7	46	7	6	10	29
4	29	7	76	7	30	10	39

Tabla 5. Comparación de los coeficientes Coseno, Jaccard y Dice contra los verdaderos relevantes.

2.4.7 Precisión y Recall

Interpolando se obtienen los resultados de Precisión y Recall para Coseno.

CON	0	10	20	30	40	50	60	70	80	90	100
1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
2	0,5	0,5	0,5	0,5	0,5	0,5	0,286	0,286	0,286	0,286	0,286
3	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25
4	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25
5	1	1	1	1	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0
7	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25
8	0	0	0	0	0	0	0	0	0	0	0
9	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	0,4	0,4	0,4	0,4	0,5	0,5	0,5
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0
18	1	1	1	1	1	1	0,2857	0,2857	0,2857	0,2857	0,2857
19	1	1	1	1	1	1	1	1	1	1	1
20	0	0	0	0	0	0	0	0	0	0	0
21	1	1	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1	1	1
23	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33
24	0,5	0,5	0,5	0,5	0,5	0,5	0,6666	0,6666	0,6666	0,6666	0,6666
25	1	1	1	1	1	1	1	1	1	1	1
26	0	0	0	0	0	0	0	0	0	0	0
27	1	1	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1	1	1
29	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0
32	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
33	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
34	0	0	0	0	0	0	0	0	0	0	0
35	1	1	1	1	1	1	1	1	1	1	1

Tabla 6. Resultados de los índices Precisión y Recall de las 35 consultas con respecto a Coseno.

La siguiente tabla muestra los resultados de la media de Precisión para Coseno, Jaccard y Dice con respecto a Recall.

	PRECISIÓN		
RECALL	Coseno	Jaccard	Dice
0	0,39970068	0,41522857	0,39215676
10	0,39970068	0,41522857	0,38571419
20	0,39970068	0,41522857	0,38571419
30	0,39970068	0,41522857	0,38571419
40	0,38255782	0,41522857	0,38571419
50	0,38255782	0,41522857	0,38571419
60	0,36078721	0,40570286	0,37619038
70	0,36078721	0,40570286	0,37619038
80	0,36364435	0,40570286	0,37619038
90	0,36364435	0,40570286	0,37619038
100	0,36364435	0,40570286	0,37619038

Tabla 7. Resultados de Precisión y Recall, respecto a los coeficientes Coseno, Jaccard y Dice.

Gráficamente se representan los resultados de Recall y la media de Precisión para Coseno, Jaccard y Dice.

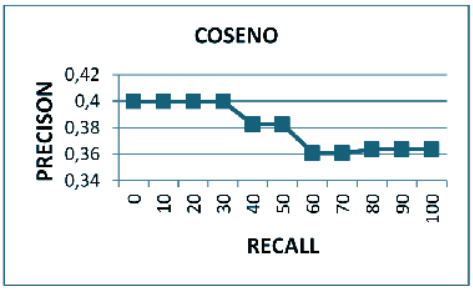

Fig. 3 Relación entre los indices Precision y Recall para el coeficiente Coseno.

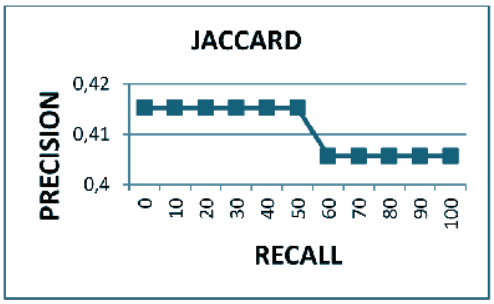

Fig. 4. Relación entre los indices Precision y Recall para el coeficiente Jaccard.

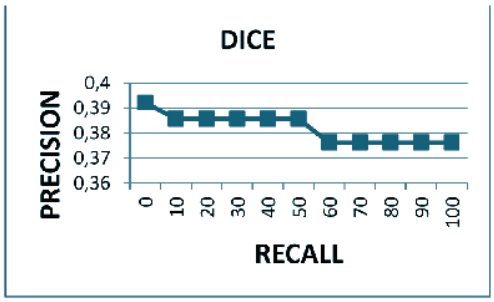

Fig. 5. Relación entre los indices Precision y Recall para el coeficiente Dice.

En las figuras 3, 4 y 5 se puede evidenciar que Precision y Recall son inversamente proporcionales, es decir, cuando Precision aumenta Recall disminuye y viceversa.

2.4.8 Determinación del factor F

El factor F, relaciona Precision y Recall (cobertura) de la siguiente forma:

Ecuación 8

$$F_{\beta=1} = \frac{\beta^2 + 1 \ * precision * recall}{\beta^2 * precision + recall}$$

El cual es una medida que combina ambos parámetros (Precision y Recall) y que inicialmente se utilizó para evaluar las prestaciones de los sistemas de recuperación de información (van Rijsbergen, 1979)[11].

Con esta medida, y variando el valor de β, se puede dar más peso a un parámetro que a otro. Si β>1, se está dando más peso a la precisión, y si β<1, a la cobertura. Normalmente se consideran ambas medidas por igual (β=1). [11]

3. RESULTADOS

En la tabla 8, se muestran los resultados obtenidos para este factor respecto a coseno, siendo β=1. De la misma forma se obtuvieron los resultados para Jaccard y Dice.

Coseno	Precision	Recall	F
Consulta 1	0,0833	0,3333	0,1333
Consulta 2	0,2857	1,0000	0,4444
Consulta 3	0,2500	0,2500	0,2500
Consulta 4	0,1429	0,5000	0,2222
Consulta 5	0,1111	0,3333	0,1667
Consulta 6	0,0000	0,0000	0,0000
Consulta 7	0,1250	0,2500	0,1667
Consulta 8	0,0000	0,0000	0,0000
Consulta 9	0,1667	0,5000	0,2500
Consulta 10	0,0000	0,0000	0,0000
Consulta 11	0,0000	0,0000	0,0000
Consulta 12	0,6000	0,6000	0,6000
Consulta 13	0,5000	0,5000	0,5000
Consulta 14	0,0000	0,0000	0,0000
Consulta 15	0,0000	0,0000	0,0000
Consulta 16	0,2000	0,3333	0,2500
Consulta 17	0,0000	0,0000	0,0000
Consulta 18	0,2857	0,6667	0,4000
Consulta 19	0,4000	0,4000	0,4000
Consulta 20	0,0000	0,0000	0,0000
Consulta 21	0,5000	0,2000	0,2857
Consulta 22	1,0000	0,1667	0,2857
Consulta 23	0,2500	1,0000	0,4000
Consulta 24	0,5000	0,2222	0,3077
Consulta 25	0,5000	0,3333	0,4000
Consulta 26	0,0000	0,0000	0,0000
Consulta 27	1,0000	0,0303	0,0588
Consulta 28	0,5000	0,2500	0,3333
Consulta 29	0,0000	0,0000	0,0000
Consulta 30	0,0000	0,0000	0,0000
Consulta 31	0,0000	0,0000	0,0000
Consulta 32	0,1429	0,5000	0,2222
Consulta 33	0,3333	0,2500	0,2857
Consulta 34	0,0000	0,0000	0,0000
Consulta 35	1,0000	0,2500	0,4000

Tabla 8. Resultado del factor F con respecto a Coseno, usando las 35 consultas.

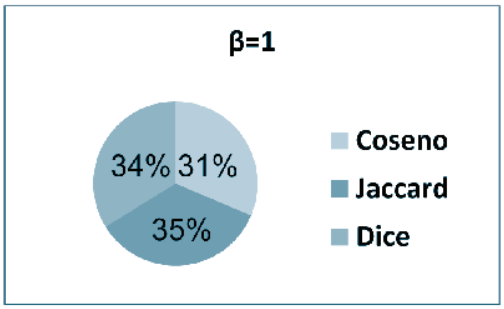

Fig. 6. Comparación del factor F entre los factores coeficientes Coseno, Jaccard y Dice para las 35 consultas, siendo β=1.

Según la figura 6 se puede determinar que el Factor F tiene mayor porcentaje en el coeficiente de Jaccard, pero esta diferencia no es significativa con respecto a los demás coeficientes.

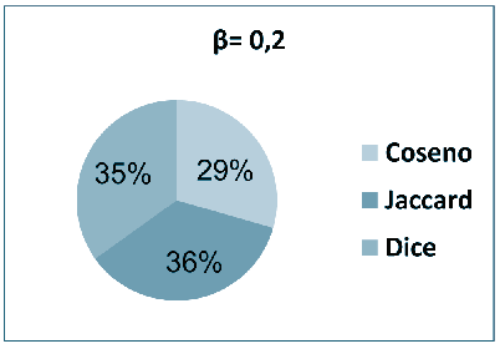

Fig. 7. Comparación del factor F entre los factores coeficientes Coseno, Jaccard y Dice para las 35 consultas, siendo β<1.

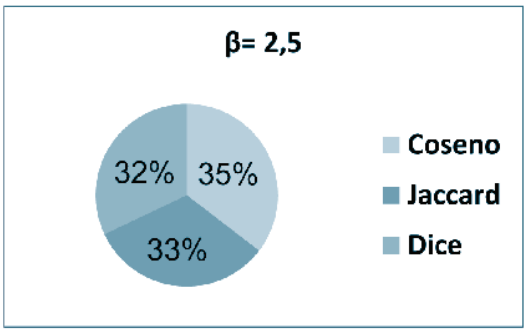

Fig. 8. Comparación del factor F entre los factores coeficientes Coseno, Jaccard y Dice para las 35 consultas, siendo β>1.

En las figuras 6, 7 y 8, se puede apreciar que el Factor F aumenta para el coeficiente Jaccard cuando β es menor que 1 y de igual manera aumenta para el coeficiente de Coseno cuando β es mayor que 1.

4. CONCLUSIONES Y FUTUROS TRABAJOS

La recuperación realizada a los archivos ADI, mostró que la mayor precisión se obtiene usando el coeficiente Jaccard.

Gráficamente se demostró que Precision y Recall son inversos y que cuando uno crece el otro disminuye al modificar los valores de β para el factor F.

El procesamiento del Lenguaje Natural con herramientas automáticas, hoy en día usa medidas de evaluación de recuperación de información, que ayudadas con una buena técnica de filtrado de información (Stop Words) ó reducción a su raíz (stemming) [13], permiten ser más precisos en la recuperación de documentos.

Para trabajos futuros se recomienda usar una base de datos con un mayor número de documentos y consultas, para verificar la consistencia de lo aquí encontrado.

REFERENCIAS

[1] Martínez, Beltrán Beatriz.Técnicas del Procesamiento del Lenguaje Natural. Puebla, México. 2007.
[2] La Serna, Nora. Roman, Ulises. Osorio, Norberto. Benito, Oscar. Espezua, Jimy. Vega, Hugo. Estudio y Evaluación de los Sistemas de Recuperación de información. 2004.
[3] Vallez, Mari y Pedraza-Jiménez, Rafael. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. http://www.hipertext.net/web/pag277.htm
[4] Jackson Peter. Moulinier Isabelle. Natural language processing for online applications: text retrieval, extraction, and categorization.
[5] Manning, C. et al. Introduction to InformationRetrieval. Cambridge University Press, 2008. URL http://www.sli.stanford.edu/~hinrich/information-retrieval-book.html
[6] Martin, P.,Sergio, Modelos de recuperación, disponible en: http://modelosderecuperacioni.iespana.es/
[7] Angel F., Rodríguez Zazo, Figuerola G., Alonso J.L. and Gómez R., Recuperación de Información utilizando el Modelo Vectorial, Departamento de informática y automática, Universidad de Salamanca, 2002, Mayo.
[8] El procesamiento del lenguaje natural, tecnología en transición. Jaime Carbonell. Congreso de la Lengua Española, Sevilla, 1992.
[9] Disponible en la página http://www.sc.ehu.es/sbweb/fisica/cursoJava/fundamentos/colecciones/stringtokenizer.htm
[10] Enríquez, Fernando, Técnicas de Bootstrapping en el Procesamiento del Lenguaje Natural.
[11] Molina, M., Antonio, Desambiguación en procesamiento del lenguajenatural mediante técnicas de aprendizaje automático.
[12] http://ir.dcs.gla.ac.uk/resources/ test_collections/
[13] Figuerola, C.G.; Gómez Díaz, R.; López de San Román, E. Stemmingand n-grams in Spanish: an evaluation of their impact on information retrieval.