Comparativa entre herramientas para la extracción de entidades espaciales geográficas

Juan Diego Gómez Fierros & Azucena Montes Rendón
Centro nacional de investigación y desarrollo tecnológico (cenidet)
Interior Internado Palmira S/N, Col. Palmira.
Cuernavaca, Morelos. México
C.P. 62490 Tel. 01(777) 362-7770
{juan.gomez10c@cenidet.edu.mx}
{amr@cenidet.edu.mx}

1. Introducción

Internet y las tecnologías de la Web 2.0 han propiciado una explosión de la información disponible en diferentes modalidades. Las técnicas de Extracción de Información son una de las alternativas para organizar y mejorar el acceso a este torrente de información. Varios métodos han sido propuestos para anotar las palabras de forma automática con las etiquetas de parte de discurso (POS, *part-of-speech*).

Algunos investigadores utilizan el sistema basado en normas [Greene and Rubin, 1971] [Brill, 1993]. Otros implementan métodos probabilísticos [Bahl and Mercer, 1976] [Church, 1988] [Cutting et al, 1992] [DeRose, 1988] [Kempe, 1993]. Finalmente, modelos de redes neuronales también se han probado en el etiquetado POS [Federic and Pirrelli, 1994] [Schmid, 1994] y los problemas relacionados de la predicción de POS [Nakamura et al, 1990].

En los últimos años han aparecido varios servicios de software comercial que permiten la extracción de palabras claves y de Entidades Nombradas (NE del inglés Named Entity) como OpenCalais[1], Zemanta[2], AlchemyAPI[3], Evri[4], STILUS Sem[5], OpenAmplify[6], SaploTags[7] o BeliefNetworks[8].

Estos servicios se han integrado en numerosas aplicaciones y es previsible que, con el avance del software como servicio, sirvan para mejorar las capacidades semánticas y de interoperabilidad de muchas más en un futuro próximo.

Algunas de ellas ya se encuentran disponibles para el procesamiento del español como Open Calais, AlchemyAPI o STILUS Sem.

2. Herramientas

En este apartado se mencionan las principales herramientas existentes para la extracción de entidades geográficas.

Solo se tomaron en cuenta herramientas que trabajan o consideran el idioma español, ya que, esta comparativa será utilizada para el desarrollo posterior de un trabajo de tesis de maestría.

2.1. OpenCalais

Open Calais es un servicio web de Thomson Reuters que permite la extracción de entidades, hechos y eventos de texto libre en inglés, francés y español. Su versión en inglés es la que presenta una mayor funcionalidad, si bien en español permite:

- reconocimiento y categorización de entidades usando 15 clases de entidades
- evaluación de la relevancia de entidades
- desambiguación y enlazado con Linked Open Data para algunos tipos como Company

Open Calais ofrece un API sencillo que puede ser usado mediante SOAP, REST vía HTTP POST, o HTTP POST. Como entrada permite documentos de distintos formatos (HTML, HTMLRAW, XML y texto).

Además de la etiquetación semántica el servicio incluye la eliminación de cabeceras y otros elementos en HTML así como la detección de idioma. Como salida ofrece la elección de varios formatos XML/RDF, texto, texto con micro formatos o JSON.

Los formatos XML/RDF y JSON incluyen URIs referénciales que pueden enlazar con una tercera fuente de conocimiento, típicamente Linked Data[9].

[1] www.opencalais.com
[2] www.zemanta.com
[3] www.alchemyapi.com
[4] www.evri.com
[5] www.daedalus.es/productos/stilus/stilus-sem
[6] www.openamplify.com
[7] www.saplo.com
[8] www.beliefnetworks.net
[9] http://linkeddata.org/

Para la definición de todas las clases utilizadas en Open Calais existe tanto un esquema RDFS[10] como una ontología OWL[11].

2.2. AlchemyAPI

AlchemyAPI utiliza la tecnología de procesamiento de lenguaje natural y algoritmos de aprendizaje automático para analizar el contenido de un texto, extracción semántica de metadatos: información sobre personas, lugares, empresas, temas, idiomas y mucho más.

Para la extracción de nombres de entidades entre las cuales se identifican a las personas, empresas, organizaciones, ciudades, lugares geográficos y otras entidades contenidas dentro de una página HTML o en un documento de texto.

Esta herramienta cuenta con un reconocimiento avanzado de nombres de entidades (NER), la capacidad de funcionar en varios idiomas y ofrece capacidades completas de desambiguación.

El etiquetado de conceptos se realiza de manera automática de forma similar a como lo realizamos los seres humanos, posee una capacidad avanzada para el marcado de concepto, el cual, es capaz de hacer abstracciones ("Hillary Clinton + Barbara Bush + Laura Bush == Primeras Damas de los Estados Unidos"), la anotación en los documentos cuenta con altos índices de exactitud.

AlchemyAPI, extrae los términos más importantes y las palabras clave "tema" de las páginas HTML y documentos de texto.

Utiliza algoritmos avanzados estadísticos y lingüística para analizar el contenido, "marcando" las palabras y las frases más importantes.

2.3. Extractiv

Extractiv ofrece dos servicios principales: Rastreo de la Semántica en páginas Web y Semántica "On-Demand". Ambos servicios ofrecen la conversión automática de los contenidos en la estructura semántica de datos, pero se diferencian en los tipos de documentos y tareas para las que son las más adecuadas.

El servicio de Rastreo de la Semántica en páginas Web, permite rastrear millones de páginas web y convertir cualquier contenido estructurado encontrado en las páginas, en los datos semánticos.

El extractor Extractiv está construido en base a una potente plataforma de distribución, que permite el proceso Extractiv más de 100.000 documentos por hora. El procesamiento del lenguaje natural (PLN) que se ejecuta con este rastreo, proporciona la extracción de información de forma precisa.

En la parte de semántica "On-Demand", se ofrece la conversión semántica automática para el procesamiento de documentos específicos, pudiendo ser aquellos documentos contenidos en la propia computadora. Utilizando el API REST "On-Demand", se pueden cargar y procesar tantos documentos como se desee.

2.4. STILUS NER

STILUS-NER es una biblioteca, que forma parte de STILUS-Core, para la detección y etiquetado de entidades con nombre (Named Entity Recognition), consiste, como su propio nombre indica, en la detección y clasificación de los elementos del texto en categorías predefinidas, como nombres de personas, organizaciones, lugares, expresiones numéricas, de tiempo, etc., que aparecen mencionadas en un texto escrito en un determinado idioma.

Actualmente están disponibles los siguientes idiomas:

- **es** - Castellano
- **it** - Italiano
- **en** - Inglés
- **fr** – Francés

3. Precisión

Este concepto fue definido por Kent [Kent et al, 1955], como factor de pertinencia. Hay otros autores que se refieren a él, como ratio de aceptación. Para Salton [Salton and McGill, 1983], la precisión es la proporción de material recuperado realmente relevante, del total de los documentos recuperados. A esta definición Frakes [Frakes and Baeza, 1992] añade que el resultado de esta operación está entre 0 y 1. Así, la recuperación perfecta es en la que únicamente se recuperan los documentos relevantes y por lo tanto tiene un valor de 1.

En esta medida, se evalúa directamente la correlación de la pregunta con la base de datos e indirectamente sirve para ver cómo es de completo el algoritmo de indización [Kowalski, 1997]. Si el algoritmo de indización tiende a generalizar teniendo un umbral alto en los términos de índice o al usar los conceptos genéricos de indización, entonces la precisión es baja, no importa cómo sea el algoritmo de similaridad entre la pregunta y el índice.

Ecuación 1 Precisión. [Salton and McGill, 1983]:

$$\text{Precisión} = \frac{Documentos\ relevantes\ recuperados}{Documentos\ recuperados}$$

[10] http://www.opencalais.com/files/RDFS%20schema_09Jun16.txt

[11] http://www.opencalais.com/files/owl.opencalais-4.3a.xml

Su representación gráfica se hace marcando en el eje de las x el número de documentos y en las de las y , los valores de precisión de 0 a 1, asociada a esos documentos recuperados de modo que los sistemas más precisos son aquellos que en su gráfica describen una curva con valores altos al principio y que van decreciendo. Comparando las distintas curvas de los sistemas, podemos hacernos una idea clara de cuáles son más precisos.

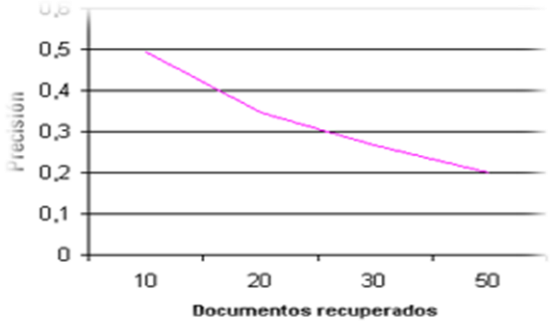


Grafico 1. Precisión

4. Exhaustividad

La exhaustividad, aunque en menor medida que la precisión, es el otro concepto más utilizado en la evaluación de los sistemas de recuperación.

Muchos autores, por influencia del término inglés la denominan " recall" o "rellamada". Es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no. Esta medida es inversamente proporcional a la precisión.

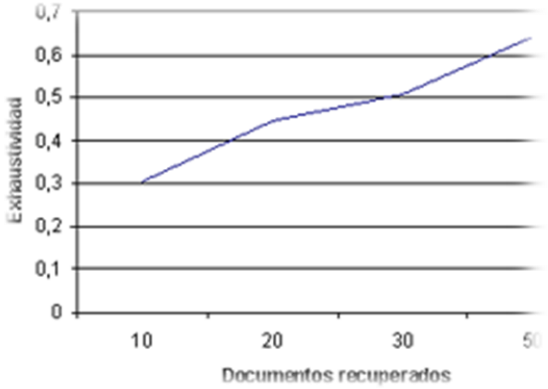
Fue formulada, al igual que la de precisión por Kent [Kent et al, 1955] con el nombre de factor de exhaustividad. Años más tarde, Swets [Swets, 1963] la llamó probabilidad condicional de un item y Goffman y Newil [Goffman and newill, 1964] la denominaron sensibilidad (sensibility).

La ecuación propuesta por [Salton and McGill, 1983]:

$$\text{Exhaustividad} = \frac{Documentos\ relevantes\ recuperados}{Documentos\ relevantes}$$

Al igual que la precisión también podemos representarla gráficamente, para ello en el eje de las x marcamos el número de documentos y en el de las y el valor de la exhaustividad calculada para cada documento. A medida que aumenta el número de documentos recuperados, recordemos que la salida es ordenada en función de la relevancia, la exhaustividad va en aumento. El comportamiento normal de esta gráfica, es que la curva vaya aumentando.

Los sistemas serán más exhaustivos cuando alcancen al principio valores altos (próximos a 1), y después vayan disminuyendo.



Grafica2. Exhaustividad

5. Relación entre precisión y exhaustividad

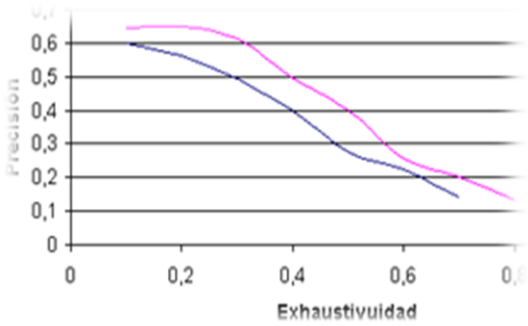
Necesitamos comprobar que la precisión y la exhaustividad están compensadas, ya que un sistema con una exhaustividad muy alta pero con baja precisión y viceversa no será adecuado. Para comprobar cómo se relacionan la precisión y la exhaustividad en una sola gráfica, podemos hacerlo de varias maneras: calculando la precisión exhaustividad interpolada: es decir tomamos un conjunto de documentos y calculamos para cada valor de precisión su exhaustividad.

Por ejemplo tomamos los veinte primeros documentos recuperados, donde hay quince documentos relevantes y calculamos la precisión y la exhaustividad para cada documento recuperado (si el primer documento recuperado es relevante tendremos una precisión de 1/1 y una exhaustividad de 1/15).

También podemos hacerlo de manera no interpolada, en este caso calculamos la exhaustividad por tramos de documentos recuperados.

Por ejemplo, tomamos veinte documentos y calculamos el valor de exhaustividad en los cinco primeros documentos recuperados, luego en los diez, luego en los quince y finalmente en los veinte documentos recuperados.

Una vez que tenemos estos valores, en ambos casos marcamos los puntos, en el eje de las x los valores correspondientes a la exhaustividad y para cada valor de ésta marcamos en el de las y el valor de la precisión que le corresponde. Uniendo los puntos obtenemos la curva que nos dice cómo se relacionan estas dos medidas en cada sistema y comparándolas ver qué sistema es el más efectivo.



Grafica 3. Precisión y exhaustividad interpoladas

[Salton and McGill, 1983], sugirieron un método para la evaluación del sistema proponiendo salidas ordenadas de los documentos en las respuestas.
De este modo, la precisión y la exhaustividad dependían del valor de corte, es decir, del punto a partir del cual se considera que al usuario ya no le interesan los documentos. Este criterio Blair lo denomina " punto de futilidad " [Blair, 1980]. La precisión y la exhaustividad se calculan para cada posición en la lista de documentos recuperados.

6. Pruebas

Se tomaron 10 noticias de diferentes periódicos en las cuales se marcaron las entidades geográficas de manera manual. Posteriormente, se introdujeron estos textos en las 4 herramientas mencionadas anteriormente y se realizaron pruebas de **precisión y exhaustividad** (*recall*), con los resultados obtenidos, se realizó una tabla comparativa para verificar la exactitud en el reconocimiento de entidades y finalmente, se muestra una gráfica en la que se observan los resultados de cada herramienta.
En la **tabla 1**, se observa en la primera columna por la izquierda, el número de noticia, en la siguiente columna el número de entidades geográficas extraídas de forma manual y en las siguientes columnas el número de entidades geográficas reconocidas por cada una de las 4 herramientas.

Número de entidades Reconocidas					
Noticia	**Manual**	**calais**	**AlchemyAPI**	**Extractiv**	**Ner**
1	3	1	1 (Tag)	1	1
2	2	1	0	2	2
3	4	0	1	2	2
4	2	2	0	3	2
5	8	5	6	3	6
6	10	7	5	8	9
7	2	2	1	0	1
8	4	4	3	4	4
9	3	4	4	2	4
10	2	2	1	1	0
Total	40	28	22	26	31

Tabla 1. Entidades reconocidas

6.1 Resultados

Para realizar los cálculos de precisión y exhaustividad, las operaciones originales fueron modificadas, ya que lo que se recuperan son entidades relevantes, no documentos relevantes. Solo se realizara una medida de precisión y otra de exhaustividad para cada una de las herramientas probadas, obteniendo así un valor global de la efectividad en estas.

Total de entidades:
Manual ("Gold Standard"): 40

CALAIS: 28
Correctas: 26
Erróneas: 2
AlchemyAPI: 22
Correctas: 20
Erróneas: 2
Extractiv: 26
Correctas: 22
Erróneas: 4
STILUS Ner: 31
Correctas: 28
Erróneas: 3

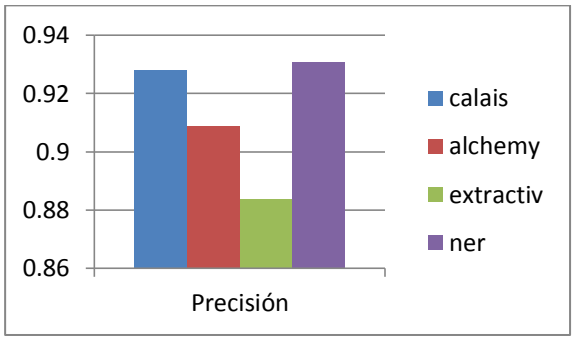
Precisión CALAIS = 26/28 = **0.928**
Exhaustividad CALAIS = 26/40 = **0.65**

Precisión AlchemyAPI = 20/22 = **0.909**
Exhaustividad AlchemyAPI= 20/40 = **0.5**

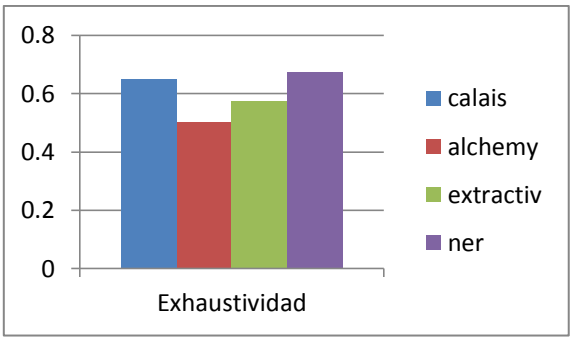
Precisión Extractiv = 22/26 = **0.846**
Exhaustividad Extractiv = 22/40 = **0.575**

Precisión Stilus ner = 28/31 = **0.931**
Exhaustividad Stilus ner = 28/40 = **0.675**

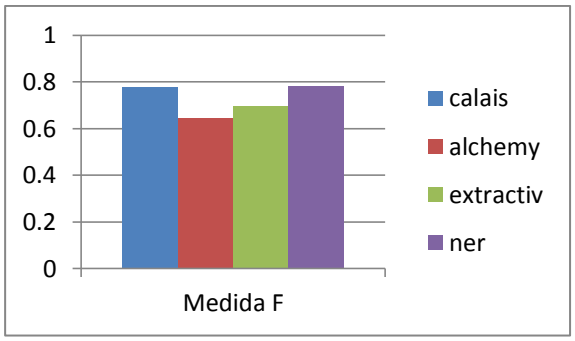
Realizando una gráfica para cada uno de los resultados anteriores:



Grafica 4. Resultados de precisión

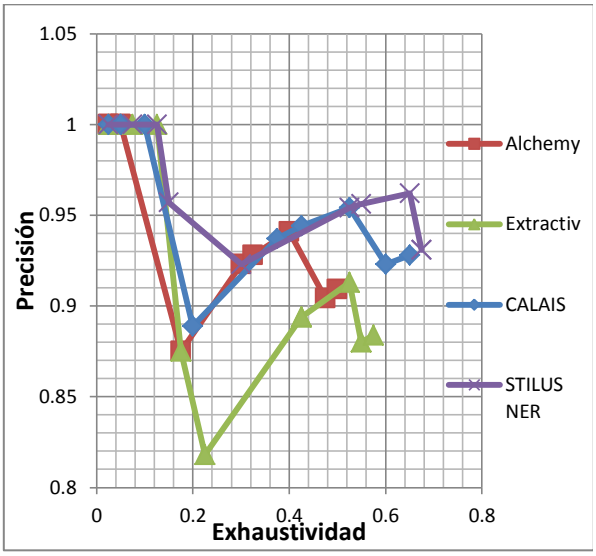


Grafica 5. Resultados de exhaustividad



Grafica 7. Resultados de medida F

En la siguiente grafica se observan las medidas de precisión y exhaustividad interpoladas:



Grafica 6. Precisión y exhaustividad interpoladas

Sólo con las medidas de precisión y exhaustividad no se puede saber cuál de las herramientas es la mejor, se necesita una nueva medida la cual nos mostrara la relación que existe entre la precisión y la exhaustividad.

Esta medida nos permitirá concluir cuál de las herramientas es la mejor opción, estadísticamente:

$$\text{Medida F} = \frac{2 * precisión * exhaustividad}{precisión + exhaustividad}$$

Medida F CALAIS = 1.2064/1.548= **0.779**
Medida F AlchemyAPI = 0.909/1.409 = **0.645**
Medida F Extractiv= 1.0166/1.459 = **0.6967**
Medida F STILUS = 1.25685/1.606 = **0.7825**

7. Conclusiones

Las primeras 3 herramientas CALAIS, AlchemyAPI y Extractiv tienen soporte multi-idioma, mostrando mejores resultados en idioma inglés, la herramienta de STILUS NERV, está diseñada para procesar documentos en español mostrando por esta razón mejores resultados.

Las herramientas de CALAIS y NERV, en sus versiones de prueba en línea, al identificar una entidad geográfica colocan, cuando es posible, la localización del lugar, por ejemplo al localizar la entidad geográfica "Cuernavaca" CALAIS coloca: Cuernavaca, Morelos, México, además de su latitud y longitud en coordenadas geográficas, por su parte NERV, coloca lo siguiente: LOCATION -> Morelos/México/América -> Cuernavaca -> Cuernavaca.

Extractiv coloca un link en su página web cuando identifica una entidad geográfica hacia dbpedia12 en donde se muestra información relevante con la entidad localizada como Coordenadas, nombre oficial, historia, extensión, población, etc.

AlchemyAPI y Open Calais cuentan con API´s para distintos lenguajes de programación de manera gratuita, para probar su funcionalidad permitiéndonos implementar las herramientas en diversos escenarios. Para poder utilizar estas API´s es necesario registrarse en las páginas y solicitar una "API KEY", que es una clave necesaria para poder tener acceso a las diferentes funcionalidades de las herramientas. También existen versiones comerciales de ambas herramientas las cuales permites mayor número de operaciones diarias y brindan soporte a los usuarios.

Las herramientas NERV y Extractiv solo cuenta con una versión comercial pero puede solicitarse una versión de prueba con utilidad reducida.

Existen más herramientas para el idioma español como las propuestas presentada por [Carreras et al, 2002], [Ferrández et al. 2006], [Florian, 2002] y [Sang, 2002], las cuales extraen entidades para el

[12] dbpedia.org

idioma español con un porcentaje de exactitud muy alto.
Finalmente se muestra en la **Tabla 2** en la que se pueden observar todas las características a tener en cuenta para elegir la herramienta a utilizar como parte del proyecto:

Herramienta	Precisión	Exhaustividad	Medida F	Libre	Comercial	API´s
STILUS Ner	0.931	0.675	0.7825	X (restringido)	X	On-line
Calais	0.928	0.65	0.779	X	X	.NET Java Ruby PHP
Extractiv	0.884	0.575	0.6967	X (restringido)	X	On-line
Alchemy	0.909	0.5	0.645	X	X	Android Java Perl Ruby Python PHP c/c++/c#

Tabla 2. Características de las herramientas

Si nos vamos por la parte de la eficiencia estadística, la mejor opción es STILUS NER, al mostrar el mayor porcentaje en Medida F y tener los valores más altos de Precisión y Exhaustividad, le sigue CALAIS quedando en segundo lugar.
En la opción de API's disponibles, ALCHEMY API es la que cuenta con el mayor número de estas y todas con versiones libres para su uso, en segundo lugar vemos a CALAIS con varios API's disponibles, además de contar con una herramienta que implementa todas las opciones de la demo en línea.
STILUS NER cuenta con una API, la cual devuelve los resultados de la extracción de entidades en un documento XML.
Tomando en cuenta estos parámetros, se decidió utilizar la herramienta CALAIS, ya que muestra un aceptable grado de precisión y exhaustividad (recall), además cuenta con varias API´S disponibles en formato libre.
También se considerara como complemento la herramienta de STILUS NER en su versión libre para las entidades espaciales más regionales, como en el caso de estados y municipios de México, provincias de España, etc.

Referencias:

[Greene and Rubin, 1971] Greene, B. and Rubin, G. (1971). Automatic grammatical tagging of english. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island.

[Brill, 1993] Brill, E. (1993). A Corpus-Based Approach To Language Learning. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

[Bahl and Mercer, 1976] Bahl, L. R. and Mercer, R. L. (1976). Part-of-speech assignment by a statistical decision algorithm. IEEE International Symposium on Information Theory, pages 88-89.

[Church, 1988] Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In Proceeding of the Second Conference on Applied Natural Language Processing, pages 136-143.

[Cutting et al, 1992] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In Proceeding of the third Conference on Applied Natural Language Processing, pages 133-140.

[DeRose, 1988] DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. Computational linguistics, 14(1).

[Kempe, 1993] Kempe, A. (1993). A probabilistic tagger and an analysis of tagging errors. Technical report, Institut für maschinelle Sprachverarbeit ung, Universität Stuttgart.

[Kent et al, 1955] Kent A. Et al. Machine literature searching. VIII. Operational Criteria for Designing Information Retrieval Systems American Documentation April 1955 6 (2) p. 93-101

[Salton and McGill, 1983] Salton, G. and M. J. McGill.. Introduction to Modern Information Retrieval. New York: McGraw Hill. 1983

[Frakes and Baeza, 1992] Frakes, W. B. and Baeza Yates, R. (ed.) Information Retrieval: data structures and Algorithms. México: Prentice-Hall, 1992

[Kowalski, 1997] Kowalski, G. Information Retrieval Systems Theory and Implementation. Boston: Kluwer Academic Publisher, 1997

[Swets, 1963] Swets, J. A. Information retrieval Systems Science, 141 (3577): July 1963 p. 245-250

[Goffman and newill, 1964] Goffman and Newill. Methodology for test and evaluation of information retrieval systems. Information Storage and Retrieval (1964) 3 p. 19-25

[Blair, 1980] Blair. Searching bases in large interactive document retrieval systems Journal of the American Society for Information Science 1980 (31) 4 p. 271-277