

Evaluación de la Selección, Traducción y Pesado de los Rasgos para la Mejora del Clustering Multilingüe

S. Montalvo¹, A. Navarro¹, R. Martínez², A. Casillas³, and V. Fresno¹

¹ Dpt. Informática, Estadística y Telemática
Universidad Rey Juan Carlos

{soto.montalvo, victor.fresno}@urjc.es, axelux@gmail.com

² Dpt. de Lenguajes y Sistemas Informáticos
UNED

raquel@lsi.uned.es

³ Dpt. Electricidad y Electrónica
Universidad del País Vasco

arantza.casillas@ehu.es

Resumen En este trabajo hemos realizado un estudio para evaluar el impacto de utilizar diferentes representaciones de documentos en el resultado del clustering multilingüe. Para ello, seguimos un modelo basado en la selección y traducción de rasgos. La selección se basa en la utilización de información sobre la categoría gramatical y el contexto. La traducción se ha llevado a cabo utilizando *EuroWordNet 1.0* y aplicando un método de desambiguación automática. Además, se han utilizado diferentes funciones de pesado de los rasgos (TF, TF-IDF y WIDF). El objetivo principal es estudiar la importancia de cada uno de estos elementos y así poder determinar una o varias combinaciones de ellos que conduzcan a obtener buenos resultados en el clustering multilingüe. La evaluación se ha llevado a cabo con un corpus comparable de noticias escritas en castellano e inglés. Se ha usado un algoritmo de clustering de partición de la librería CLUTO y la calidad de los resultados se ha determinado mediante una medida de evaluación externa. Los mejores resultados se obtienen representando con las entidades nombradas de todo el documento y con las funciones de pesado TF y TF-IDF.

1. Introducción

El clustering multilingüe parte de un conjunto de documentos escritos en varios idiomas y tiene como objetivo agruparlos de manera que se puedan obtener clusters o grupos multilingües. Un cluster multilingüe contendrá aquellos documentos que estén relacionados o traten del mismo tema aunque estén escritos en diferentes lenguas. Mientras que un cluster monolingüe estará compuesto únicamente de documentos relacionados escritos en el mismo idioma.

El aumento de la cantidad de documentos electrónicos escritos en diferentes lenguas conlleva la necesidad de desarrollar sistemas que manejen toda esta

información y faciliten su acceso a los potenciales usuarios. El clustering multilingüe puede facilitar tareas como la recuperación de información multilingüe (agrupando documentos antes y después de la recuperación), alineación de corpora paralelos y comparables, entrenamiento de parámetros para sistemas de traducción automática estadística, etc.

Los diferentes enfoques a la hora de abordar el clustering multilingüe se pueden clasificar en dos grandes grupos: por un lado, aquellos que hacen uso de técnicas de traducción y, por otro, aquellos que transforman el documento en una representación independiente del lenguaje.

En los sistemas basados en traducción, bien para traducir los documentos completos a una lengua eje, o bien para seleccionar ciertos rasgos y sólo traducir éstos a una lengua eje, es crucial la exactitud de la traducción obtenida. Los recursos bilingües que se utilizan, normalmente ofrecen varias posibilidades o sentidos en la traducción de una palabra y no es trivial elegir el adecuado. Aunque se pueden aplicar métodos de desambiguación automática, éstos no están libres de errores y no elegir el sentido de la traducción apropiado puede conducir a un agrupamiento erróneo.

Por otro lado, los sistemas que transforman el documento en una representación independiente del lenguaje tienen ciertas limitaciones. Por ejemplo, aquellos que trabajan con tesauros dependen fundamentalmente del alcance de éstos. La identificación de datos numéricos y de fechas puede resultar muy apropiada para ciertos tipos de clustering y documentos. Sin embargo, en otros casos este tipo de datos puede no ser relevante y resultar una fuente de ruido. Además, la identificación de cognados, que suele ser otra forma de representación independiente, está muy ligada al tipo de lenguas de que conste el corpus.

En este trabajo presentamos los resultados de un estudio que hemos llevado a cabo para evaluar el impacto de la utilización de diferentes representaciones de los documentos en el resultado del clustering multilingüe. Para ello, hemos utilizado el modelo basado en la selección y traducción de los rasgos. La selección se ha basado en seleccionar o no rasgos pertenecientes a diferentes categorías gramaticales, entidades nombradas y determinados contextos. La traducción se ha llevado a cabo mediante *EuroWordNet 1.0* [Vossen 1998], aplicando un método de desambiguación automática. También hemos utilizado diferentes funciones de pesado de los rasgos (TF, TF-IDF y WIDF). El objetivo principal es estudiar la importancia de cada uno de estos aspectos y, así, poder determinar una o varias combinaciones de ellos que conduzcan a la obtención de buenos resultados en el clustering multilingüe.

La evaluación se ha llevado a cabo con un corpus comparable de noticias escritas en castellano e inglés. Con el fin de utilizar medidas de evaluación externa, se ha recopilado un subconjunto de noticias comparable que ha sido agrupado manualmente y que ha servido como solución de referencia. Como el énfasis del estudio se ha puesto en la selección y traducción de los rasgos y no en el algoritmo de clustering, se ha utilizado un algoritmo de partición bien conocido en la literatura.

El resto del artículo se estructura como sigue: en la Sección 2 se describen brevemente algunos trabajos relacionados. La Sección 3 explica, mostrando cada una de sus fases, el estudio realizado. En la Sección 4 se presenta la colección utilizada en la evaluación, así como los experimentos junto con los resultados obtenidos. Por último, la Sección 5 incluye las conclusiones y trabajo futuro.

2. Trabajos Relacionados

El clustering multilingüe de documentos normalmente se aplica sobre corpus paralelos [Silva et. al. 2004] o corpus comparables ([Rauber et. al. 2001], [Mathieu et. al 2004], [Pouliquen et. al. 2004], [Chen and Lin 2000], [Steinberger et. al. 2002], [Lawrence 2003]).

Si tenemos en cuenta los trabajos basados en el uso de técnicas de traducción, se emplean dos estrategias: (1) traducir el documento completo a una lengua eje, y (2) traducir algunos rasgos del documento a una lengua eje.

Con respecto a la primera aproximación, algunos autores utilizan sistemas de traducción automática, mientras que otros traducen el documento palabra a palabra, consultando un diccionario bilingüe. En [Lawrence 2003] se presentan varios experimentos de clustering sobre un corpus comparable de Ruso e Inglés; varios de estos experimentos están basados en el uso de sistemas de traducción automática.

Cuando se trata de traducir sólo algunos rasgos del documento, en primer lugar es necesario seleccionar qué rasgos se van a traducir (normalmente nombres, entidades nombradas, verbos y adjetivos) para, a continuación, traducir dichos rasgos mediante un diccionario bilingüe o consultando un corpus paralelo.

En [Mathieu et. al 2004], antes del proceso de clustering, se lleva a cabo un análisis lingüístico que extrae los lemas y reconoce entidades nombradas de diversas categorías (lugar, organización, persona, expresión temporal, expresión numérica, evento). Por lo tanto, los documentos se representan mediante un conjunto de rasgos. Además, los autores tienen en cuenta su frecuencia para seleccionar los rasgos más relevantes. Finalmente, utilizan un diccionario bilingüe para traducir los rasgos seleccionados. En [Rauber et. al. 2001] los autores presentan una metodología que consiste en la extracción de todas las palabras que aparecen en n documentos, exceptuando las palabras vacías de contenido. Posteriormente, mediante sistemas de traducción automática construyen un corpus monolingüe. Una vez finalizado el proceso de traducción, de forma automática, los documentos se organizan en diferentes clusters usando un método de aprendizaje no supervisado mediante redes neuronales.

Algunas aproximaciones llevan a cabo un proceso de clustering independiente en los documentos de cada lengua, es decir, un clustering monolingüe. Posteriormente, tratan de encontrar relaciones entre los clusters monolingües obtenidos, generando así clusters multilingües. Sin embargo, otros trabajos comienzan con un proceso de clustering multilingüe buscando relaciones entre los documentos de todos los idiomas involucrados. Este es el caso de [Chen and Lin 2000], donde los autores proponen una arquitectura de resúmenes

de noticias que incluye un proceso de clustering monolingüe y multilingüe. El clustering multilingüe toma como entrada los resultados de una fase previa de clustering monolingüe. Los autores seleccionan diferentes tipos de rasgos dependiendo del tipo de clustering: para el clustering monolingüe usan entidades nombradas y para el clustering multilingüe, además, tienen en cuenta los verbos.

Las estrategias de clustering que generan para cada documento una representación independiente del idioma en el que está escrito, intentan estandarizar o normalizar los contenidos de varias formas: (1) mapeando los contenidos a una representación independiente, o (2) reconociendo rasgos independientes del lenguaje dentro del texto. Ambas posibilidades se pueden emplear de forma aislada o combinada.

La primera aproximación requiere la existencia de recursos lingüísticos multilingües, como tesauros, para crear una representación del texto que consista en un conjunto de entradas de tesoro. Normalmente, en un tesoro multilingüe, los elementos de las diferentes lenguas se relacionan mediante entradas independientes de la lengua. Por lo tanto, dos documentos escritos en distinto idioma pueden ser considerados similares si tienen una representación parecida de acuerdo a lo que indica el tesoro. En algunos casos, es necesario el uso de tesauros combinados con métodos de aprendizaje automático para realizar un mapeo correcto de los documentos con el tesoro. En [Steinberger et. al. 2002] calculan la similitud semántica representando los contenidos de los documentos de forma independiente a la lengua en la que están escritos, por medio del tesoro *Eurovoc*.

La segunda aproximación, reconocer en el texto rasgos independientes de la lengua, implica poder identificar elementos como: fechas, números y entidades nombradas. Por ejemplo, en [Silva et. al. 2004] los autores presentan un método basado en lo que denominan Expresiones Relevantes (ER). Una expresión relevante es una unidad léxica de cualquier longitud extraída de los documentos mediante la herramienta LiPXtractor. Las expresiones relevantes se usan para extraer un conjunto de rasgos, pero los clusters obtenidos son monolingües.

Otros trabajos combinan la identificación de rasgos independientes del idioma (como números, fechas, ...) con el mapeo de los rasgos del texto con un tesoro. En [Pouliquen et. al. 2004] la similitud entre los clusters multilingües se basa en la combinación lineal de tres tipos de entradas: (a) cognados, (b) detección automática de nombres de referencias geográficas, y (c) los resultados de un proceso de mapeo de un sistema de clasificación multilingüe, que mapea los documentos en un tesoro multilingüe (*Eurovoc*).

En [Steinberger et. al. 2004] proponen extraer características independientes del idioma usando gazettters y expresiones regulares, además de tesauros y sistemas de clasificación.

3. Representación y clustering de documentos

Para la representación de los documentos utilizamos el modelo de espacio vectorial [Salton and McGill 1983]. Según este modelo, para cada documento se

obtiene un vector en el que cada componente representa el peso de un rasgo en dicho documento.

Nuestra propuesta para el clustering multilingüe de documentos se compone de las siguientes fases:

1. Selección de rasgos.
2. Generación de la representación intermedia.
3. Traducción de rasgos.
4. Generación de la representación final.
5. Clustering.

3.1. Selección de rasgos

En esta primera fase se seleccionan los rasgos que se van a tener en cuenta en la representación de cada documento. En nuestro enfoque, esta selección requiere que el corpus esté analizado morfo-sintácticamente, lematizado y con las entidades nombradas identificadas y categorizadas. En este trabajo únicamente hemos tenido en cuenta las entidades de las categorías PERSONA, LUGAR, ORGANIZACIÓN y MISCELÁNEA.

La selección se basa en 3 aspectos:

- La categoría gramatical.
Normalmente se consideran categorías más discriminantes los nombres, verbos y adjetivos.
- Ser o no entidad nombrada.
En el caso particular de los documentos de noticias, tiene sentido que las entidades nombradas sean consideradas como rasgos realmente discriminantes. Por ello, en principio, cabría pensar que las representaciones que incluyan dicho tipo de rasgos conseguirán mejores resultados en el clustering que aquellas que no los consideren. Aunque realmente serán útiles para la representación en la medida en que los recursos utilizados para la traducción sean capaces de dar cuenta de ellas.
- El contexto.
Otro aspecto a tener en cuenta en la selección de rasgos es el contexto en el que se encuentran. En el estilo periodístico es habitual que en el primer párrafo se resuma el contenido primordial de la noticia. De ahí que nosotros hayamos considerado dos tipos de contexto en nuestro estudio: el documento completo y el primer párrafo.

3.2. Generación de la representación intermedia

Una vez extraídos los rasgos que se van a tener en cuenta, se generará una representación intermedia por cada parte monolingüe del corpus.

Esta representación intermedia consiste en una matriz, donde cada fila es un vector que se corresponde con uno de los documentos del corpus. Cada columna representa uno de los rasgos que aparecen en el corpus y que ha sido seleccionado.

Los valores de cada componente de los diferentes vectores que forman la matriz se asignan por medio de funciones de peso. En este estudio hemos utilizado funciones bien conocidas en la literatura (TF, TF-IDF y WIDF) que se describen a continuación:

Term Frequency, TF [Luhn 1957]: cada rasgo tiene una importancia proporcional al número de veces que aparece en el documento.

Inverse Term Frequency, TF-IDF [Salton and Yang 1973]: la combinación de pesos de un término t en un documento d , siendo N el número de documentos y $df(t)$ el número de documentos que contienen el rasgo t , viene dada por:

$$TF - IDF(d, t) = TF(d, t) \times IDF(t); \quad IDF(t) = \log \frac{N}{df(t)} \quad (1)$$

Weighted Inverse Term Frequency, WIDF [Salton 1989]: extensión de IDF que incorpora la frecuencia del término sobre la colección de documentos:

$$WIDF(d, t) = TF(d, t) \sum_{i \in D} TF(i, t) \quad (2)$$

3.3. Traducción de rasgos

Para traducir los rasgos se dispone de la base de datos léxica *EuroWordNet 1.0*. Con este recurso se traducen a castellano todos los rasgos que aparecen en la representación intermedia del corpus en inglés.

En la traducción, uno de los factores clave es la desambiguación automática. Cuando para un rasgo en inglés se obtiene más de un sentido posible como traducción hemos aplicado el método de desambiguación que se describe a continuación. De los diferentes sentidos obtenidos por *EuroWordNet*, se elige aquél que esté presente entre los rasgos de la matriz del corpus en castellano. Nuestra hipótesis es que dado que trabajamos con un corpus comparable, esperamos que la traducción correcta de una palabra aparezca, en la mayoría de los casos, en el corpus del otro idioma.

En el caso de que al intentar traducir un rasgo no se encuentre ninguna traducción, dicho rasgo se elimina de la representación, salvo que se trate de una entidad nombrada. Así, se contempla que pueda haber entidades nombradas, aunque no se puedan traducir, ya que pueden coincidir en ambas lenguas.

3.4. Generación de representación final

Una vez que se han generado dos representaciones intermedias, una por cada corpus monolingüe y, además, la del corpus en inglés se ha traducido, se fusionan en una única representación. Entonces, como representación final para el proceso de clustering multilingüe se dispone de una única matriz.

3.5. Clustering

Para realizar el clustering usamos un algoritmo de partición. En particular, el algoritmo *Direct* de la conocida librería CLUTO [Karypis 2002]. El número total de clusters que se quieren obtener es un dato que hay que proporcionar al algoritmo.

4. Evaluación

En esta sección se presenta el corpus con el que se realiza la evaluación, así como los experimentos realizados y los resultados obtenidos.

4.1. Corpus

Un corpus comparable es una colección de textos similares en diferentes idiomas o diferentes variedades de un mismo idioma. En este trabajo usamos una colección de noticias escritas en inglés y castellano, referentes al mismo periodo de tiempo. Las noticias se encuentran clasificadas y se trata de noticias de la agencia EFE que han sido recopiladas en el proyecto HERMES⁴. Esta colección se puede considerar como un corpus comparable. Para realizar la evaluación hemos usado un subconjunto de noticias que se compone de 79 noticias en castellano y 70 noticias en inglés, en total 149 noticias.

Para poder comprobar la bondad de los resultados del algoritmo de clustering con las diferentes representaciones se ha realizado una agrupación manual de la colección. Tres personas han sido las encargadas de leer los documentos y formar grupos atendiendo a sus contenidos. La solución manual se compone de 26 clusters, siendo todos ellos multilingües.

4.2. Experimentos y Resultados

Se realizaron experimentos con diferentes combinaciones de todos los criterios de selección de rasgos descritos en la Sección 3.

La calidad de los resultados se ha evaluado mediante una medida de evaluación externa, la medida-F [van Rijsbergen 1974]. Esta medida compara la solución obtenida por nuestro sistema con la solución humana. La medida-F combina las medidas de precisión y recall:

$$F(i, j) = \frac{2 \times \text{Recall}(i, j) \times \text{Precision}(i, j)}{(\text{Precision}(i, j) + \text{Recall}(i, j))}, \quad (3)$$

denominamos clase al grupo de la solución humana y cluster al grupo devuelto por el sistema, así: $\text{Recall}(i, j) = \frac{n_{ij}}{n_i}$, $\text{Precision}(i, j) = \frac{n_{ij}}{n_j}$, donde n_{ij} es el

⁴ <http://nlp.uned.es/hermes/index.html>

número de miembros de la clase i en el cluster j , n_j es el número de miembros del cluster j y n_i el número de miembros de la clase i . Para todos los clusters:

$$F = \sum_i \frac{n_i}{n} \max\{F(i, j)\}, \quad (4)$$

donde n es el número de documentos. Esta función está acotada entre los valores 0 y 1, que representan la peor y mejor calidad de clustering respectivamente.

En la tabla 1 se presentan los mejores resultados obtenidos con las diferentes medidas de pesado utilizadas, TF, TF-IDF y WIDF respectivamente.

La primera columna de las tablas indica la categoría gramatical de los rasgos seleccionados: NOM (nombres), VER (verbos), ADJ (adjetivos), NE (entidades nombradas) y 1^{er} PAR (todos los rasgos de las categorías seleccionadas que aparezcan en el primer párrafo). La segunda columna representa la medida-F y la tercera columna indica la relación entre el número de clusters multilingües obtenidos y el número total de clusters multilingües que se deberían haber obtenido. Recuérdese que en el corpus de evaluación la solución manual tenía todos los clusters multilingües.

Tabla 1. Resultados de clustering con las diversas representaciones

Rasgos seleccionados	F. peso	medida-F	Clusters Multl./Total
NOM, VER	TF	0.8164	16/26
NOM, VER, 1 ^{er} PAR	TF	0.7214	15/26
NOM, ADJ	TF	0.8555	18/26
NOM, ADJ, 1 ^{er} PAR	TF	0.7769	21/26
NOM, VER, ADJ	TF	0.8027	16/26
NOM, VER, ADJ, 1 ^{er} PAR	TF	0.7321	14/26
NE	TF	0.8628	18/26
NE, 1 ^{er} PAR	TF	0.7012	15/26
NOM, VER	TF-IDF	0.8534	21/26
NOM, VER, 1 ^{er} PAR	TF-IDF	0.7372	19/26
NOM, ADJ	TF-IDF	0.8406	21/26
NOM, ADJ, 1 ^{er} PAR	TF-IDF	0.7517	22/26
NOM, VER, ADJ	TF-IDF	0.7984	20/26
NOM, VER, ADJ, 1 ^{er} PAR	TF-IDF	0.7570	21/26
NE	TF-IDF	0.8117	19/26
NE, 1 ^{er} PAR	TF-IDF	0.6823	21/26
NOM, VER	WIDF	0.6705	26/26
NOM, VER, 1 ^{er} PAR	WIDF	0.5560	25/26
NOM, ADJ	WIDF	0.7302	26/26
NOM, ADJ, 1 ^{er} PAR	WIDF	0.6486	26/26
NOM, VER, ADJ	WIDF	0.7090	26/26
NOM, VER, ADJ, 1 ^{er} PAR	WIDF	0.6155	25/26
NE	WIDF	0.7323	24/26
NE, 1 ^{er} PAR	WIDF	0.6747	22/26

Como era de esperar los resultados varían en función de la representación utilizada.

Los mejores valores de la medida-F se obtienen, en general, con las funciones de pesado TF y TF-IDF, quedando a bastante distancia las representaciones con la función WIDF. El único punto a favor de esta última es que sus soluciones, aunque de peor calidad, obtienen un número de clusters multilingües más cercano al de la solución manual.

En cuanto al tipo de rasgos, las entidades nombradas (NE) obtienen los mejores valores de medida-F en dos de las representaciones, y el tercer mejor valor en la otra. Estos resultados indican que se trata de rasgos muy representativos de los documentos del corpus. Por otra parte, las representaciones con NOM, ADJ y VER también obtienen buenos resultados con las tres funciones de pesado.

5. Conclusiones y Trabajos Futuros

Hemos realizado un estudio para determinar el impacto de la utilización de diferentes representaciones de los documentos en el resultado del clustering multilingüe. Para ello, partiendo del modelo basado en la traducción de rasgos de los documentos, el énfasis se ha puesto en una selección de rasgos basada en información obtenida de: las categorías gramaticales, el uso de las entidades nombradas y la elección del contexto. Además, se han utilizado diferentes funciones de pesado de los rasgos. Para la desambiguación en la traducción de los rasgos se ha propuesto un método sencillo basado en la naturaleza de los corpus comparables.

La experimentación se ha realizado sobre un corpus comparable de noticias escritas en castellano e inglés y se ha utilizado un conocido algoritmo de clustering.

Los resultados indican que las representaciones obtenidas con las funciones de pesado TF y TF-IDF obtienen un agrupamiento de más calidad que con la función WIDF. Por otra parte, las entidades nombradas (NE) resultan ser los rasgos que mejor representan el corpus utilizado en la experimentación, un corpus de noticias. También las representaciones con rasgos de tipo NOM, ADJ y VER muestran un buen comportamiento.

En cuanto a la elección del contexto, se aprecia que a menor tamaño de contexto los resultados empeoran. En todas las representaciones cuyo contexto es el primer párrafo, los resultados empeoran con respecto a las representaciones cuyo contexto es el documento completo.

Como posibles trabajos futuros están combinar diferentes recursos para la traducción con el fin de aumentar el número de entidades nombradas que se traducen, por ejemplo utilizar gazettters además de la base de datos léxica y aplicar reglas de equivalencia. Asimismo, la distinción entre las diferentes categorías de entidades nombradas puede resultar útil en la fase de selección de rasgos.

Referencias

- [Karypis 2002] Karypis G.: "CLUTO: A Clustering Toolkit". Technical Report: 02-017. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455, 2002.
- [Mathieu et. al 2004] Benoit Mathieu and Romanic Besancon and Christian Fluhr. "Multilingual document clusters discovery". RIAO 2004, p. 1-10, 2004.
- [Pouliquen et. al. 2004] Bruno Pouliquen and Ralf Steinberger and Camelia Ignat and Emilia Käsper and Irina Temikova. "Multilingual and cross-lingual news topic tracking". Proceedings of the 20th International Conference on computational Linguistics, p. 23-27, 2004.
- [Rauber et. al. 2001] Andreas Rauber and Michael Dittenbach and Dieter Merkl. "Towards Automatic Content-Based Organization of Multilingual Digital Libraries: An English, French, and German View of the Russian Information Agency Novosti News". Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections Petrozavodsk, RCDI'2001.
- [Silva et. al. 2004] Joaquin Silva and J. Mexia and Carlos Coelho and Gabriel Lopes. "A Statistical Approach for Multilingual Document Clustering and Topic Extraction form Clusters". Pliska Studia Mathematica Bulgarica, v.16, p. 207-228, 2004.
- [Vossen 1998] Vossen, P. "Introduction to EuroWordNet". Computers and the Humanities Special Issue on EuroWordNet, 1998.
- [Chen and Lin 2000] Hsin-Hsi Chen and Chuan-Jie Lin. "A Multilingual News Summarizer". Proceedings of 18th International Conference on Computational Linguistics, p. 159-165, 2000.
- [Steinberger et. al. 2002] Ralf Steinberger and Bruno Pouliquen and Johan Scheer. "Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC". CICling'2002, p. 415-424.
- [Lawrence 2003] Lawrence J. Leftin. "Newsblaster Russian-English Clustering Performance Analysis". Columbia computer science Technical Reports.
- [Steinberger et. al. 2004] Ralf Steinberger and Bruno Pouliquen and Camelia Ignat. "Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications". SILTC 2004.
- [Luhn 1957] H. P. Luhn. "A statistical approach to mechanized encoding and searching of literaty information". IBM Journal of Research and Development, 1957.
- [Salton and McGill 1983] G. Salton and M. McHill. "Introduction to Modern Information Retrieval". McGraw-Hill, New York. 1983.
- [Salton and Yang 1973] G. Salton and C.S. Yang. "On the specification of term values in automatic indexing". Journal of Documentation, 1973.
- [Salton 1989] G. Salton. "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer". Addison-Wesley, 1989.
- [van Rijsbergen 1974] C.J. van Rijsbergen. "Foundations of evaluation". Journal of Documentation, 30, p. 365-373, 1974.
- [Vossen 1998] P. Vossen. "Introduction to EuroWordNet". Computers and the Humanities Special Issue on EuroWordNet. 1998.