Information Retrieval Based on Text Extraction*

Jiménez-Salazar, H. Pinto-Avendaño, D. & Salazar-Martínez, H.

Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
hjimenez@fcfm.buap.mx, dpinto@cs.buap.mx, hilariocam@yahoo.com.mx

Abstract. The work presented here describes an experiment carried out to explore a semantic representation of texts. We use a corpus as a basis of the representation of the words by its sense relationship. Thus, a text is represented by an extract which applies a method that selects the most representative sentences. In the retrieval process, the queries are expanded using the same corpus that supports the extraction method, and then overlapped with the text representation in order to know the relevance of the text. The experiment was performed using Spanish language texts and shows the behavior, into the boolean model, of three expansion techniques: using EuroWordNet, using a raw corpus in order to expand queries and the use of mutual information to select some words of expanded queries. The last technique achieved the better results.
Keywords: Semantic text representation, Text extraction, Query expansion.

1 Introduction

Semantic representation of text has became a great challenge because it inherits the main characteristics of natural language: ambiguity, irregularity and universality [7]. Such characteristics impact most of the Information Retrieval Systems (IRS) issues such as the user's query at the web ("specific", "broad" and "vague" queries [1]). The imprecision of actual IRS may be due, in essence, to the following statement: *if some query terms are contained in a document, then said document is relevant to the query, to some degree.* However, a more natural statement may in fact be construed, such as the following: *if some query terms are related to the conceptual content of a document, then said document is relevant to the query.* Query expansion techniques try to follow the second statement which focuses on the concept of the query and has shown a performance improvement [20], [14]. In our view, it is necessary to explore innovative approaches to deal with the second statement.

In many applications of natural language processing, the selection of features of a text has been a central topic. Since Salton's proposal [17], the use of terms such as index has dominated the applications and development of IRS. Indeed,

* This work was partially supported by Mexican CoNaCyT grant Nr. I39165-A, and SEP-SESIC-DGES-VIEP grant Nr. III09G02.

Bhanu Prasad (Editor): IICAI-03, pp. 443-454, 2003.

such an approach to text representation coincides with methods that make use of terms instead concepts.

In order to manage concepts we need operations such as abstraction and extension. This idea comes from formal concept theory, where a concept has two parts: intent (features) and extent (instances) [3]. In this work, we approximate the abstraction operation as the extract of a text and the query expansion as an extension of query's words. So, the extract of a text is treated as an alternative way to represent textual information. This representation is a part of the management of a text, but it is also necessary take into account the context in which a document is presented. We are using a corpus as the knowledge source of both text extraction and query expansion. Certainly, there are some linguistic resources that may support the above approach. Nowadays, however, it is difficult to achieve resources for different domains, in contrast with the vast quantity of texts that may be used. Therefore, the use of a corpus as a linguistic resource may be an advantage.

The next section states the background about the representation using sense relationships. Sections 3 and 4 explain some details of the text representation and query expansion techniques used here, respectively. The next two sections deal with how the resources are employed, as well as the description of the experiment. At the end, the conclusions of the present work are given.

2 Sense Relationship

Sense relationships of a word x is the set of the whole of words that have a relation with x [10]. We conceived sense relationships approximated by first association terms, which have been used in some works with promising results [15], [5], [19]. Several approaches have been carried out to approximate a semantic representation of a text [6], [11]. Some of them are expensive because of the use of large linguistic resources [16], [18].

We used the most representative sentences of a document in order to represent every document, and we also used sense relationships to expand queries in order to improve recall and precision when words belonging to the query did not appear in any document to be searched.

Our approach is based on formal concept theory [3]. In formal concept theory each concept is represented by two parts, the intent (i.e. the set of features of the concept), and the extent (the set of instances that have such features). More formally, let G be a set of instances, M a set of features and ψ a correspondence form G to M. The set $\{m \in M | \psi(g) = m\}$ is denoted by $\Psi(g), g \in G$. A concept in $\langle G, M, \psi \rangle$ is a pair (A, B), where $A \subset G$, $B \subset M$, and it holds

$$\bigcap_{a \in A} \Psi(a) = B. \tag{1}$$

There are some applications of formal concepts to textual representation regarding IRS [13] and the approximation to some lexical relationships [8] based on the order defined between concepts.

Formal concept theory is very elegant but impractical in most cases, if it is not applied adequately. For example, we note that an instance x may exist which shares several features with the intent of A, in a concept (A, B), but that x has no features in B. Then, x is not an instance of (A, B), which is unnatural.

Because a sentence is a full semantic unit, we may take a sentence as an instance in formal concept theory. Through syntax and semantics, sense relationships of each word are combined to achieve the sense of a sentence [4], although this combination is not completely known. At text representation, sentences may be seen as instances and sense relationships as features. This leads to the definition of the formal concept "the use of the word x" as

$$(\{S_1, \ldots, S_k\}, \{x\}), \tag{2}$$

where $S_1, \ldots, S_k$ are sentences that contain x. Taking S_i as a set, according to (1) we have $\cap_i S_i = x$, then the features as sense relationships of x related to its use do not appear in the intent of the concept. Here we see how some features are discarded due to the rigid application of formal concept theory.

We therefore use formal concept theory as an approximation to a natural phenomena, and with the present work we try to understand the behavior of sense relationships in text representation.

3 Text Representation

In order to represent a document, we extract the most representative sentences using an inexpensive method proposed at [9] that takes into account the semantic content of texts. The method is supported by the idea of sense relationships attached to a word. Certainly we use an approximation of sense relationships of a word x as the whole of words in sentences where x occurs.

Our approach uses only a raw corpus as the linguistic resource. The method is very simple: first, we represent each of the sentences with the contexts of its components (words), then we find out the similarity between a sentence z and the whole text except by z (the complement of z). The more similarity between a sentence z and its complement, the more representative z is of the text.

The procedure to represent a document has three steps: preprocessing, text expansion and text extraction.

3.1 Preprocessing

The preprocessing stage is applied to both the input text T and the corpus used C. The goal of this stage is to identify every sentence that makes up the text and its stemmed words. The algorithm is simple and does each of the following four steps sequentially: tokens identification, stopwords elimination, stemming, and sentence segmentation. After applying each step to the text T as well to the corpus, the text obtained is called T_1 and C_1 respectively.

3.2 Text Expansion

The sense relationships with respect to the corpus C_1 is a set of pairs:

$$V = \{(x, y) | x, y \in S, \text{ for a sentence } S \in C_1\}. \tag{3}$$

Let us represent the expansion of a word x as

$$\text{Expand}(x, V) = \{y | (x, y) \in V\}. \tag{4}$$

The expansion of a word is seen as a set of words (without repetitions). From the expansion of a word, we can expand a sentence making up the corresponding tuple to S, $(x_1, \ldots, x_k)$:

$$\text{ExpandS}(S, V) = (\text{Expand}(x_1, V), \ldots, \text{Expand}(x_k, V)). \tag{5}$$

The expansion T_2 of the text T_1 is then:

$$T_2 = (\text{ExpandS}(S_1, V), \ldots, \text{ExpandS}(S_n, V)). \tag{6}$$

where $T_1 = (S_1, S_2, \ldots, S_n)$ and n is the number of sentences. The complexity of the text expansion process is $O(k)$, where k is the number of words of the text to be summarized.

3.3 Text Extraction

This module uses a similarity function in order to sort every sentence S_i from T_2 in decreasing order, based on their similarity with the same text T_2. The similarity function used here is a simplification of the Jaccard similarity function, which is defined as:

$$sim(S_i, \bar{S}_i) = \#(S_i \cap \bar{S}_i). \tag{7}$$

sim counts the number of common elements of S_i and $\bar{S}_i$, where S_i denotes the i-th sentence of T_2 and $\bar{S}_i$ the complement of S_i in T_2 (i.e. T_2 without S_i). The next specification describes the sentence ranking:

$$T_3 = permut(T, \pi_1 \circ sort(i, sim(S_i, \bar{S}_i))_i) \tag{8}$$

where π_1 projects the first argument of the tuples, $permut$ arranges T according to its second argument, and $T_2 = (S_i)_i$.

In order to select the most representative sentences of a text, it is necessary to find a threshold. The definition of such a value must be researched in order to determine the best extract for each application.

4 Query Expansion

The query expansion method is performed by two ways, by using EuroWordNet and by using a corpus to provide sense relationships of each preprocessed

query word, as we saw in the above section. For EuroWordNet a query word is expanded using those words that are linked with some of the following relations: *has-holo-madeof, has-holo-member, has-holo-part, has-hyperonym, has-hyponym , has-mero-madeof, has-mero-member, has-mero-part, is-caused-by, role, role-agent, role-instrument, role-location, role-patient, xpos-fuzzynym and xpos-near-synonym.* Let us denote any relation of the previous set as R_{EW}. Thus a word x is expanded by the following set:

$$\mathrm{Expand}_{\mathrm{EW}}(x) = \{y | x R_{\mathrm{EW}} y\}. \tag{9}$$

Corpus-based expansion was implemented taking the whole of the sense relationships of a query word from the corpus:

$$\mathrm{Expand}(x, V) = \{y | (x, y) \in V\}, \tag{10}$$

as well as using the mutual information regarding a threshold of 6[1]:

$$\mathrm{Expand'}(x, V) = \{y | (x, y) \in V \& MI(x, y) \geq 6\}, \tag{11}$$

where $MI(x, y)$ is the mutual information of x and y:

$$MI(x, y) = log_2 \left(\frac{N \cdot fr(x, y)}{fr(x) \cdot fr(y)} + 1 \right), \tag{12}$$

using $fr(x)$ as the number of sentences of C_1 which contain x and N the number of words of C_1.

5 Experiments

In our experiments, we used a corpus to define a sense relationship between terms of the query and the terms that already exist in the corpus. The corpus has defined subjects that allow us to approximately determine the sense relationships for our application. The compiled corpus is described below, as well as the dataset, queries used and the experimental results.

5.1 Corpus

The compiled corpus is composed of 96 processed documents, and related to the following subjects: politics, education, religion, economy, justice, culture, government, society, science and technology. Certainly we could have taken a corpus from a determined domain, but in this experiment we wanted to know the behavior of the representation in general, as well the influence of a heterogeneous corpus on this task. The percentage of every subject compared to the number of documents by subject can be seen in table (1). The vocabulary has 22,201 stemmed terms and 4,294 sentences.

[1] This threshold was suggested in [2].

Subject	# of docs. by subject	%
Justice	10	10.40
Culture	7	7.30
Politics	25	26.00
Society	23	24.00
Government	9	9.40
Science	11	11.50
Technology	1	1.00
Religion	2	2.10
Economy	8	8.30

Table 1. Percentage by subject of every document of the compiled corpus.

5.2 Dataset

In order to verify the performance of the algorithm, we used 103 documents with different subjects, 97 of them related with the subjects used at the corpus. A better description of dataset subject distribution is given in table (2).

Subject	Education	Economy	Culture	Justice	Politics	Society	Sports
# of documents	4	13	17	12	44	11	2
% of documents	3.8	12.6	16.5	11.6	42.7	10.6	1.9

Table 2. Dataset subject distribution.

5.3 Queries

We wrote five queries from the original documents, using words that in most cases did not exist in the representation of documents. Every word of each query was expanded by its related words, using EuroWordNet and using sense relationship from the compiled corpus. We made sure that the queries were related to the topics we were using. As an example, we show the extract of a document used in the test[2].

> *el asalto contra las 20 cárceles, efectuado durante la madrugada, tenía como objetivo detener la huelga de hambre que realizaban unos 1100 reos de extrema izquierda, 61 días antes, los cuales protestaban por su traslado a otras cárceles con celdas más pequeñas, para 3 reos como máximo, en lugar de los actuales pabellones hasta para 100 internos, y donde los izquierdistas y los grupos islámicos extremistas habían instalado centros de adoctrinamiento.*

[2] This text, as well as the whole corpus, was taken from the Mexican newspaper **La Jornada** (2000).

The objective of the assault against the 20 prisons, carried out during the dawn, was to stop the hunger strike being carried out by about 1100 extreme leftist criminals 61 days earlier, who were protesting their transfer to other jails with smaller cells for 3 criminals at the most, instead of the present pavilions for 100 prisoners, and where the extremist leftists and Islamic groups had installed indoctrination centers.

The above text is relevant to the query: *Desacuerdo de normas impuestas a reclusos (Disagreement of norms imposed to inmates).*

The set of queries used in the test is shown in table (3).

Query Number	Query
19	otorgan crédito a la nación (grant credit to the nation)
21	nombramiento de autoridad en organismo de investigación en el país (leadership appointment in research organism in the country)
30	proporción baja de egreso de estudiantes en estudios posteriores a la profesión (lower proportion of graduating students, according to subsequent occupational studies)
70	desacuerdo de normas impuestas a los reclusos (disagreement over norms imposed on prisoners)
106	delincuentes legalizan dinero sucio bancario (criminals legalize illicit bank money)

Table 3. Set of queries used at the test.

Queries are natural language sentences; furthermore, they are truncated or lemmatized and without stopwords. Thus, it is not always correct to translate queries into boolean expressions, eventhough we are now using a disjunction operation between their components.

There were five queries, each of which corresponded to one document of dataset. In this manner we hoped to get these documents as an answer from the system after submitting such queries. The results of this experiment are described in the next section.

6 Experimental Results

The three query expansion techniques described above will be referred to as:

EW All words contained in the thesaurus, using eq. (9).

Corpus All words taken from the contexts of the corpus, eq. (10).

MI Words from the corpus with high MI, eq. (11).

With the purpose of having an approximation about the documents retrieved, we show the first three documents (see table (4)) retrieved after submitting

query number 19. This query refers to the economy subject (see table (3)). We can see in table (4) that in all rows (expansion techniques) exist at least one document from the economy subject. Tables (7) and (8), show the extract of every document of table (4).

Expansion	Docs retrieved
None	124, 187, 146
Corpus	114, 30, 19
EW	124, 187, 114
MI	19, 124, 186

Table 4. First three documents retrieved using query 19.

Table (5) shows the complete results of the three types of query expansion given using the same set of text extract documents. Table (5) contains two results: the position of the correct document in the ranked list and the number of documents in the retrieved list. The ranking procedure uses the notion of membership degree to a set, used in fuzzy and rough set theory [12]: Given a set S and an element x, the membership degree of x in S is

$$\mu_S(x) = \frac{\#(S \cap x)}{\#x}, \tag{13}$$

provided that x is represented by sense relationships. Relevance of a document D to a query q is defined as $\mu_q(D)$. The values in table (5) were determined on the basis of the ranking calculated with $\mu_q(D)$.

Expansion/Query	A	B	C	D	E
None	∞/10	10/13	∞/7	∞/5	23/38
EW	∞/28	44/54	10/40	5/58	19/69
Corpus	3/98	56/103	5/103	1/56	9/103
MI	1/36	1/30	∞/30	∞/27	5/5

Table 5. Document position and number of retrieved documents.

For the particular case of only one relevant document for each query, the average precision was calculated using the formulae: $\frac{1}{p \times m}$, where p is the position and m is the number of retrieved documents shown to in table (5). This is displayed in table (6).

7 Discussion

We have presented a proposal to represent documents retrieval by means of different query expansion techniques. The IR system based on this method uses

Expansion/Query	A	B	C	D	E
None	0	0.0769	0	0	0.00114
EW	0	0.00042	0.0025	0.00344	0.00076
Corpus	0.0034	0.00017	0.00194	0.01785	0.00107
MI	0.0277	0.0333	0	0	0.04

Table 6. Average precision of query expansion.

a representation of texts based on their extracts. In this process we used the expanded query for the set of text extracts, in order to retrieve the original texts.

The experiment performed shows better results for **MI** expansion. In table (5), we can observe the following facts:

- **None**. The test displays documents that have words in common in the documents of the dataset (**B** and **D**). Besides, the number of retrieved documents indicates that there are several documents with various words of the queries. So, the word occurrence in a document does not entail documents related to concepts included in the queries.
- **EW**. This expansion only enhances the success of **None** in the position of the document, whereas the number of retrieved documents increases.
- **Corpus**. This expansion gives a very good position to all queries except for query **B**. The number of retrieved documents, however, is the whole dataset (except for queries **A** and **D**).
- **MI**. A great rise of position is given by this expansion. Using **MI**, cutting out the less important terms of expansion from the corpus leads to discarding the results of the **Corpus** test for queries **C** and **D**. In addition, the number of documents is drastically reduced.

It is a remarkable fact that the precision of the query expansion, using **MI** improves the simple use of the query expansion using the **Corpus** (except for queries **C** and **D**). In general, **MI** has better results, as we can see in table (6).

In our experiments, we generate extracts of documents with a threshold of 2. This mean that every extract has exactly 2 sentences from the original document. We would like to evaluate different thresholds (between 1 and 5) in more detail, in order to analyze the influence of those extracts on the retrieval.

We manually built five queries related to five documents of dataset. Every query was constructed carefully, so that the words belonging to each query do not appear in the related document. The most important results are that even though in most cases every word of a query does not appear in its related document, we obtained said documents, and in some cases they appeared in first place in the retrieval results.

One of our perspectives is to experiment with several corpora, varying the size, domain and dataset that belong entirely to said domain. In this case we think it is possible for the problem become more complex, due to the fact that all the documents of the dataset may compete with every user query.

# of doc	Extract
124	*Bajo los efectos de la debacle financiera de diciembre de 1994 se instaura una tercera oleada privatizadora que incluye a las principales empresas públicas de la nación.* *La vulnerabilidad financiera abrió las puertas de par en par a la inversión extranjera directa y de portafolio.*
114	*Constituiría un importante avance en el proceso de consolidación de la democracia siempre y cuando la entendamos como una forma de gobierno en la cual los ciudadanos puedan intervenir cada vez más en la toma de decisiones sobre los grandes problemas nacionales y no como una simple forma de dominio.* *Otro más fue el referéndum hecho en países como Francia para adherirse al Tratado de Maastricht y uno más el que se realizará en Irlanda después de la firma de un acuerdo entre el gobierno británico y sus opositores.*
187	*Si queremos al país como un conjunto diferenciado pero bien soldado como nación y patria podría decirse que la oposición de esa y de las otras siglas también gobierna al México plural desde los niveles estatales legislativos y de los municipios.* *Es entre todos y con las discrepancias que se debatan cómo el país habrá de resolver sus a veces insoportables problemas actuales y los que vengan partiendo del respeto recíproco y de la función específica que los votos asignan.*
30	*El titular del área que define los posgrados de excelencia –nivel muy disputado por las instituciones educativas, porque eso les permite recibir a alumnos becados por el gobierno– admitió que en este sexenio se abrieron sólo dos centros de investigación, por lo que el gobierno es en parte responsable, aunque consideró que las universidades también tienen problemas.* *En un balance de las estrategias del gobierno zedillista para mejorar la calidad de los posgrados, Martuscelli sealó que las universidades privadas se interesan poco en formar recursos de alto nivel, pues de 460 posgrados de excelencia, las instituciones particulares apenas concentran 7 por ciento de maestrías y 3 por ciento de doctorados.*
146	*En su prepotencia Lott olvida que el Presidente de México representa a la nación mexicana y que el solo intento de reconvenir a un jefe de Estado lo ubica como una persona insolente y provocadora .* *En el ámbito internacional México se ha afanado entre otras cosas por 1 sustentar su política exterior en principios 2 diversificar sus relaciones 3 participar activamente en foros internacionales y 4 privilegiar negociaciones multilaterales.*
19	*Los programas de apoyo a México del Banco Mundial buscarán expandir los programas de desarrollo rural, tratarán los problemas estructurales en el sector energía y mejorarán la protección ambiental.* *Los administradores buscarán altos rendimientos en todo el mundo, y México está bien posicionado para atraer cuantiosa parte de este capital.*
187	*Este hablista faramallero aparente jefe superior del gobierno del estado de Guanajuato y obligado vecino de la ciudad de Santa Fe y Real de Minas tecnicazo en venta de refrescos publicista sin comparación don Vicente Fox cree en sus sueos de leonés opulento que de veras va la gente a donde va Vicente.* *Aquí en el Rastro la semana pasada los marchantes le dijeron que lo conocían en su casa no allí y pienso en los miles de mercados de mi tierra que necesitan su visita para salir y seguir adelante en la sequía de los devastados campos ayer granero hoy todavía esperando los bordos prometidos por otros locuaces que el alharaquear no empobrece.*

Table 7. Text extracted from some documents (in Spanish).

# of doc	Extract
124	*(Under the effects of the financial disaster of December, 1994, a third privatizing wave was established that includes the nations main public enterprises.)* *(The financial vulnerability opened its doors to direct foreign and portfolio investment.)*
114	*(It would be an important advance in the process of the consolidation of democracy, if and only if we understand it as a form of government in which citizens could intervene in the decisions about big national problems, and not only as a simple way of domination.)* *(One more was the referendum made by countries such as France in order to adhere to the Maastricht Treaty, and another one will be made in Ireland after signing an agreement between the British government and its opposition.)*
187	*(If we want the country to be like a differentiated but well-consolidated nation and homeland, it could be said that the opposition of that and other centuries also governs the plural Mexico from its state and county legislative levels.)* *(They debate with each other and with the discrepancies about how the country will have to resolve both its current, almost unbearable problems, along with those yet to come, based on reciprocal respect and on the specific function assigned by the votes.)*
30	*(The holder of the area that defines postdegree studies of excellence – level very disputed by the educative institutions, because that allows them to receive to students granted a scholarship by the government – admitted that in this six months only two research centers were opened, reason why the government is partly responsible, although he considered that the universities also have problems.)* *(In a balance of the strategies of the zedillista government to improve the quality of postdegree studies, Martuscelli indicated that the private universities are few interested on forming resources of high level, because of from 460 postdegree studies of excellence, the particular institutions only concentrate 7 percents of master degree and 3 percents of doctorship degree.)*
146	*(In his arrogance, Lott forgets that the president of Mexico represents the Mexican nation, and that the mere attempt of reprimanding a head of state would condemn him as an insolent and provocative person.)* *(In the international environment, México has been working, among other things, on: 1. Basing its external politics on principles,2. Diversifying its relationships, 3. Actively participating in international forums, and 4. Giving priority to multilateral negotiations.)*
19	*(The programs from support to Mexico of the World Bank will look for to expand the programs of rural development, they will treat the structural problems in the sector energy and they will improve the environmental protection.)* *(Administrators will look for anywhere in the world for high performances, and Mexico is well positioned to attract numerous part of this capital.)*
186	*(This loud-mouthed speaker, apparent head of the Guanajuato state government, obligated resident of Santa Fe - Real de Minas City, skilled sodas sales technician and incomparable publicist, Mr. Vicente Fox believes in his dreams of opulent lions, in which the people really go where Vicente goes.)* *(Here in El Rastro, last week the protesters told him that they met him at his house, not there, so I think about thousands of markets from my country that need his visit in order to improve the drought of the devastated fields; yesterday (Im) a farmer, today (Im) still for the waiting for the things promised by other crazy people, that the empty promises wont fulfill.)*

Table 8. Text extracted from some documents (translated from table (7)).

References

1. Baeza-Yates, Ricardo: Searching the world wide web; challenges and partial solutions, Progress in Artificial intelligence, Coelho, Helder (Ed.), IBERAMIA98, *Lecture Notes in Artificial Intelligence*, 1484, Springer Verlag, pp 39-51, 1998.
2. Church, K.W. & Hanks P.: Word association norms, mutual information and lexicography, *Computational Linguistics*, 16 (1) pp 22-29, 1990.
3. Ganter, B. & Wille, R.: *Formal Concept Analysis*, Springer, 1999.
4. Garcia, J.F.: Estructura conceptual y comunicación, *Dimensión antropológica*, Año 2, Vol. 3, pp. 75-84, México, 1995.
5. Grefenstette, Gregory: Automatic thesaurus generation from raw text using knowledge-poor techniques, *Xerox Grenoble Lab. Report*, 1995.
6. Gauch, Susan & Wang, Jianying: A corpus approach for automatic query expansion, *Proc. of 6th Int. Conf. on Information Knowledge Management*, pp 278-284, 1997.
7. Hajičova, Eva: Procesamiento de lenguaje natural, 1er. curso internacional de sistemas expertos, CINVESTAV, Martínez-Enríquez, Ana Ma. (Ed.), México, pp 113-26, 1987.
8. Jiménez-Salazar, H.: A method of automatic detection of lexical relationships using a raw corpus, *Lecture Notes in Computer Science* A. Gelbukh (Ed.) 2588, pp 325-328, 2003.
9. Jiménez-Salazar, H.; Salazar-Martínez, H. & Pinto-Avendaño, D.: Text extraction: a sense based approach, to appear, 2003.
10. Lyons, J.: *Semantics*, Cambridge University Press, 1977.
11. Mano, Hiroko & Ogawa, Yasushi: Selecting expansion terms in automatic query expansion, *Proc of the 24th ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pp 390-391, 2001.
12. Pawlak, Z.: "Rough sets" in *Int. J. Computer and Information Science*, V.11, Poland, pp 341-65, 1982.
13. Pedersen, G.S.: A browser for bibliographic information retrieval based on an application of lattice theory, *ACM-SIGIR*, pp 270-279, 1993.
14. Qiu, Yonggang & Frei, H.P.: Concept based query expansion, *Proc of the 16th ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pp 160-169, USA, 1993.
15. Ruge, Gerda: Combining corpus linguistics and human memory models for automatic term association, *Text information retrieval*, T. Strzalkowski (Ed.), Kluwer, 1999.
16. Saggion, Horacio; Pastra, Katerina & Wilks, Yorick: Using natural language processing for semantic indexing of scene-of-crime photographs, *Lecture Notes in Computer Science* A. Gelbukh (Ed.) 2588, pp 526-536, 2003.
17. Salton, G.; Yang, C.S. & Yu, C.T.: A theory of term importance in automatic text analysis, *Journal of American Society for Information Science*, 26(1), 33-44, 1975.
18. Tiun, Sabrina; Abdullah, Rosni & Kong, Tang: Automatic topic identification using ontology hierarchy, *Lecture Notes in Computer Science* A. Gelbukh (Ed.) 2004, pp 444-453, 2001.
19. Varaschin, Gasperin and Strube De Lima: Experiment on extracting semantic relations from syntactic relations, *Lecture Notes in Computer Science* A. Gelbukh (Ed.) 2588, 2003.
20. Xu, J. & W.B. Croft: Query expansion using local and global document analysis *Proc. ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pp 4-11, Zurich, 1996.