A. Gelbukh. *Text mining: new techniques and applications*. Technical report, Serie Azul, N 34, August 1999. CIC, IPN, ISBN 970-18-3512-3, 18 pp..

Text Mining:
New Techniques and Applications

Manuel Montes-y-Gómez,
Aurelio López-López,
Alexander F. Gelbukh,
Grigori Sidorov,
Adolfo Guzmán-Arenas

Preface

This paper collection consists of the works presented at Text Mining Workshop at IJCAI-99, the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 31 July - 6 August 1999. It consists of three individual chapters:

Manuel Montes-y-Gómez, Alexander F. Gelbukh, Aurelio López-López.
Text Mining as a Social Thermometer

Manuel Montes-y-Gómez, Alexander F. Gelbukh, Aurelio López-López.
Extraction of Document Intentions from Titles

Alexander Gelbukh, Grigori Sidorov, Adolfo Guzmán-Arenas.
Text Categorization Using a Hierarchical Topic Dictionary

In the first chapter, a novel application of text mining techniques in sociology is presented. In the second chapter, a technique of text mining based on the extraction of the information from the document titles is described. In the third chapter, the use of a hierarchical dictionary for text mining tasks is discussed.

The work was accomplished in the Natural Language Laboratory of the Center for Computing Research of the National Polytechnic Institute, Mexico City, and in the INAOE, Puebla.

Text Mining as a Social Thermometer

Manuel Montes-y-Gómez,
Alexander F. Gelbukh

mmontesg@susu.inaoep.mx
gelbukh@pollux.cic.ipn.mx
Natural Language Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN).
Av. Juan Dios Bátiz, Zacatenco, 07738 D.F.
Mexico

Aurelio López-López

allopez@inaoep.mx
Electronics, INAOE
Luis Enrique Erro No. 1
Tonatzintla Pue. 72840
México

Abstract

In this paper, we show how text mining techniques can be used in analysis of Internet and newspaper news. We present a method that focuses on the current topics of opinion appearing in the news, illustrating the method mostly with Spanish examples. This method uses a classical statistical model based on distribution analysis, average calculus, and standard deviation computation to discover information on how society interests are changing and in which direction this change points. We also describe a method to identify important current topics of opinion, those that lead to stability within a period.

1 Introduction[*]

"Data Mining and Knowledge Discovery address the needs of alphanumeric databases. Text Mining is directed at textbases. The implications are that the equivalent of Knowledge Discovery is Undiscovered Public Knowledge. If this is true, this work could be the most important effort underway today" [tryb.org site, 1998].

Without a doubt, newspapers and Internet news remain one of the most important information media that reflect most current social interests. This is why we consider interesting and useful to apply text mining techniques on them.

The principal aim of our system is to analyze news and to discover the main opinion topics, their trends and some description patterns. We consider opinions to be especially important for investigating the state of society and related sociological and political issues. Indeed, opinions are not determined so directly by the interests and intentions of the columnists and professional

[*] Work partially funded by CONACyT (doctoral stydy scholarship to M. Montes-y-Gómez), CONACyT (26424A) and REDII-CONACyT projects, and SNI (grant to A. Gelbukh), Mexico.

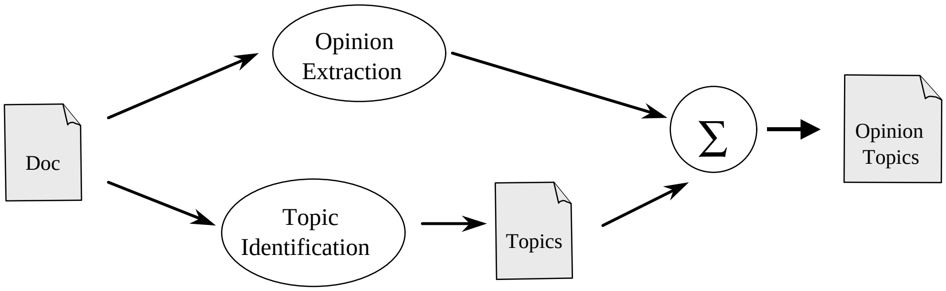


Figure 1. Module for extraction of opinion topics.

writers; instead, they represent more or less directly the *vox populi,* thus allowing to identificate the topics that are important for ordinary people.

Other systems similar to our that are focused to the analysis of some document collections has been developed [Feldman and Dagan, 1995; Lent *et al.*, 1997]. However, the work with opinions appearing in the news has some specifics. For example, it faces a double problem: (1) the discovery of changing trends and (2) the identification of states characterizing the periods of stability.

2 Source information

All Text Mining systems face the problem of obtaining the input information, or, in other words, the problem of making a structure out of the raw texts to be analyzed. This situation has caused many of the existing text mining systems to work over easy-to-extract information such as document keywords, themes or topics, proper names, or other types of simple strings [Feldman and Dagan, 1995].

Considering this problem and the way others have resolved it, we designed an opinion acquisition method based on well-known indexing and information extraction techniques. Figure 1 shows the architecture of our opinion topic extraction system.

The system consists of three modules. The first module finds the topic(s) of the document using a method similar to that proposed by [Gay and Croft, 1990], when the topics are related to noun strings.

The second module extracts the opinion paragraphs basing on so-called pattern matching technique [Kitani *et al.*, 1994] using as a trigger a list of verbs denoting communication actions, such as Spanish *dijo* ‘say’, *propuso* ‘propose’, etc. [Klavans and Kan, 1998].

The third module matches topics with opinion paragraphs and selects only the topics that are explicitly mentioned in these opinions. This new set of topics is what we call the opinion-topic set.

After this process, the input data (opinion-topic set) is complemented with the information about the opinion subject, e.g., *economía* ‘economics’, *política* ‘politics’, *sociales* ‘social’, etc. An example of the output of the process described above which is used as the input for our text mining component is shown in the Figure 2.

3 Trend analysis

After the opinion topics have been extracted from a set of news texts, the mining process begins to analyze these topics with the aim of finding and characterizing their trends. The opinion topic trend analysis has two main parts:

- Trend discovery,
- Identification of the factors (opinion topics) that contribute to produce this trend.

It also considers two different situations: *trends of change* and *stability trends*. In case of a change trend, it is important to discover the main change sources, for instance, the opinion topics with the maximum change rates. In case of stability trends, it is important to identify the stability factors, for instance, those of the most discussed opinion topics that remained without change.

3.1 Discovering trends

We discover trends in our opinion topic database comparing probability distributions [Glymour *et al.*, 1997]. These distributions have been used before for the same purpose [Feldman and Dagan, 1995], but with a different similarity measure, e.g. Feldman and Dagan used the relative entropy measure (KL-distance).

To determine the changes, we fix two “time moments,” one “past” and another “current” moment, and compare the characteristics of the two data sets, the “past” and the “current” one. We use some area values to compare the past frequency distribution D_1 with the current, or last, data distribution D_2, where the distributions D_1 and D_2 are first integrated and filtered to describe the same and relevant opinion topics.

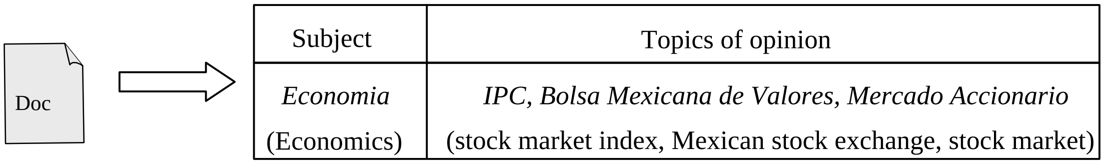
Let T_1, T_2 be sets of opinion topics at the times t_1 and t_2 respectively, with $t_1 < t_2$, and f_i^1, f_i^2 be the frequencies of the opinion topics (q_i) at the times t_1 and t_2.

<u>Integration:</u>

$$T = T_1 \cup T_2$$

This operation ensures that the work is done at the same data sets, even if some of the topics appearing at the moment t_1 have disappeared at t_2 and vice versa.

<u>Filtering</u>:



Subject	Topics of opinion
Economia (Economics)	*IPC, Bolsa Mexicana de Valores, Mercado Accionario* (stock market index, Mexican stock exchange, stock market)

Figure 2. Example of Opinion Topics to be used for Mining.

$$T' = \left\{ q_i \in T \mid f_i^1 + f_i^2 > b \right\}$$

This action removes those opinion topics, which are irrelevant to the analysis, for the sake of comutational simplicity. The frequency threshold value β specifies the minimum total frequency for a topic q_i to be considered as interesting.

Based on this set, the frequencies and probabilities can be recalculated as follows:

$$f'^{\,k}_{i} = \begin{cases} f_i^k & if\ q_i \in T_k \\ 0 & otherwise \end{cases}$$

$$p'^{\,k}_{i} = \frac{f'^{\,k}_{i}}{\sum_{j=1}^{n} f'^{\,k}_{j}}, \quad k \in \{1, 2\}$$

Comparison method:

Our purpose is to compare two probability distributions D_1 and D_2 to discover whether these two distributions are different or similar. To obtain a measure of the relation of these distributions, we compare the two areas: the change area and the maximal area. Figure 3 shows a simple example of two distributions, their change area, and their maximal area.

Change area:

$$A_c = \sum_{i=1}^{n} \left| p'^{\,1}_{i} - p'^{\,2}_{i} \right|$$

Maximal area:

$$A_m = \sum_{i=1}^{n} max\left(p'^{\,1}_{i}, p'^{\,2}_{i} \right)$$

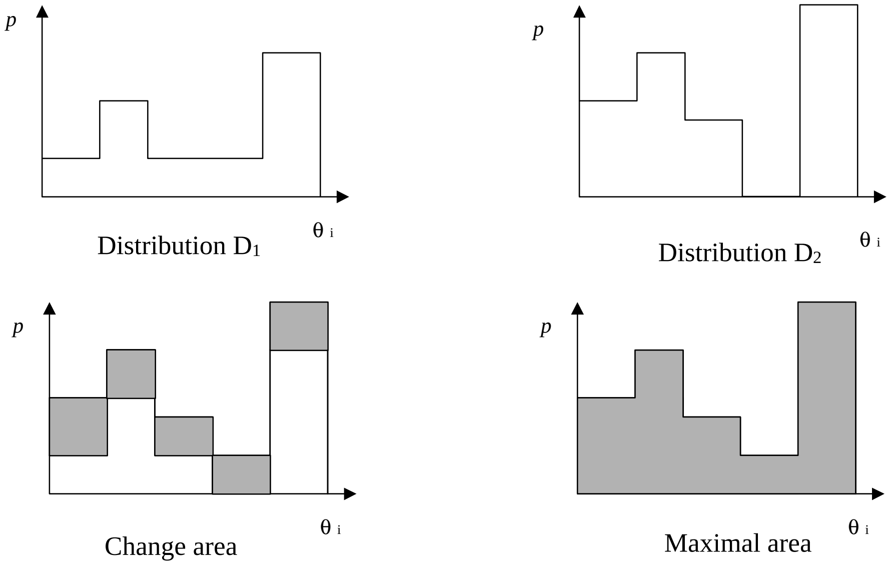
Coefficient of relation:

$$C_c = \frac{A_c}{A_m}$$

The trend discovery criteria:

- If $C_c >> 0.5$ then there exists a global change trend;
- if $C_c << 0.5$ then there exists a stability period.

If the coefficient differs from 0.5, there exists a global change trend, slighter or greater, or a stability period.



Distribution D1

Distribution D2

Change area

Maximal area

Figure 3. Comparison method.

3.2 Identification of Change Factors

A global trend of change is caused basically by abrupt changes of individual opinion-topics, that is, their identification in most cases makes evident the causes of this global change behavior.

We calculate a frequency difference value (*dF*) for each of the opinion-topics, and then select as a change factors the topics with the highest *dF* value.

$$\forall q_i \in T':$$

$$p_i^k = \frac{f'^{\,k}_i}{\sum_j f'^{\,k}_j} \qquad k \in \{1,2\} \qquad \textit{Frequency value of topic } q_i$$

$$dF_i = p_i^2 - p_i^1 \qquad \textit{Frequency difference value}$$

$$dF_m = \frac{\sum dF_i}{N} \qquad \textit{Difference average}$$

$$dF_s = \sqrt{\frac{\sum \left(dF_i - dF_m\right)^2}{N}} \qquad \textit{Difference desviation}$$

So, the opinion-topic q_i is a change factor if:

$$dF_i < dF_\mu - dF_\sigma \quad \text{or} \quad dF_i > dF_\mu + dF_\sigma$$

This criterion was found empirically. In other fields (or with other topics) the Chebyshev criteria may do as well.

3.3 Stability Factors

In general terms, stability is produced by all topics, but the most important topics are those contributing more significantly to produce this trend. The criterion we are using to identify the stability factors is as follows.

Selection of important topics:

$$T'' = \{q_i \in T' \mid f'^{1}_{i} >> f'^{1}_{m} \text{ and } f'^{2}_{i} >> f'^{2}_{m}\}$$

$$\text{where} \quad f'^{k}_{m} = \frac{\sum f'^{k}_{i}}{|T'|} \quad \text{with} \quad k \in \{1,2\}$$

$$\text{Stability factors}: \ (SF) = \{q_i \in T'' \mid q_i \in T_1 \cap T_2\}$$

4 Experimental results

To test these ideas, we analyzed *El Universal*, a Mexican newspaper, and collected the economic news for the last week of January 1999 and for the first week of February 1999. Before normalization, we had:

$$| T_{t1} \cup T_{t2} | = 47 \text{ opinion topics.}$$

After integration and filtering:

$$| T' | = 15 \text{ opinion topics, where}$$

$$T' = \{q_i \mid f'^{1}_{i} + f'^{2}_{i} > b = 1\}$$

For each opinion-topic in *T'*, we calculated its frequency (*f'*), probability (*p'*) and a difference-frequency value (*dF*). Table 1 shows these statistics.

Trend Discovery:

$$\mathrm{A_c} = 1.017$$

$$\mathrm{A_m} = 1.566$$

$$\mathbf{C_r} = \mathbf{0.65}$$

Since $C_r > 0.5$, there exists a slight global change trend.

Factors of change:

Since $dF_\mu = 2 \times 10^{-4}$ and $dF_\sigma = 0.08311$ for the opinion topic set, the change factors discovered are:

- Opinion topics that are disappearing: ($dF_i < dF_m - dF_s$):
  bancos ‘banks’,
  meta inflacionaria ‘inflationary goal’,
  inflación ‘inflation’.
- Opinion topics that are becoming more interesting: ($dF_i > dF_m + dF_s$):
  tasa de intereses ‘interest rate’,
  Brasil,
  cambio de moneda ‘change of currency’.

Topics	f´ 1	f´ 2	p´ 1	p´ 2	dF
Bancos ‘banks’	7	4	.212	.125	–.087
Meta inflacionaria ‘inflationary goal’	3	0	.09	.0	–.09
Política monetaria ‘Monetary policy’	4	4	.121	.125	.004
Ajuste fiscal ‘Fiscal adjustment’	2	0	.06	.0	–.06
Inflación ‘inflation’	4	0	.121	.0	–.121
Union monetaria ‘Monetary union’	2	0	.06	.0	–.06
Tasa de intereses ‘interest rate’	3	9	.09	.28	.19
Política fiscal ‘fiscal policy’	2	0	.06	.0	–.06
Economías asiaticas ‘Asian Economies’	1	1	.03	.031	.001
Brasil ‘Brasil’	1	4	.03	.125	.095
Economía nacional ‘national economy’	2	1	.06	.031	–.029
Cambio de moneda ‘change of currency’	0	3	.0	.094	.094
Mercado accionario ‘stock market’	2	2	.06	.062	.002
Crisis financiera ‘Financial crisis’	0	2	.0	.062	.062
Mercados financieros ‘Financial market’	0	2	.0	.062	.062

Table 1. Experimental results.

5 Conclusions and future work

These experiments and results encourage us to continue working in this direction. We have shown that it is possible to obtain useful information from not very complex text representation, though we believe that robust text representations can improve the system and will allow the design of more sophisticated text mining tools, such as inference tools, relational processes, clustering methods, visualization techniques, and summarization. As further work, we plan to:

- Enrich the topics beyond keywords, with the aim of handling themes generalizing single words. Namely, we plan to test the resources proposed in [Guzmán, 1998]. Their use will allow generalizing or specializing the topics for different levels of analysis.
- Develop a method to discover the change relations between opinion areas. For example: How the topic of the Soccer World Cup, or a general increment in sport topics, affect the general trend of the political topics?
- Analyze and classify the opinions on types. For example, opinions in which something is proposed predicted or qualified. This classification could be interesting and useful for a high level analysis of opinions.

Like with any data mining system, the more data and data types we have the more, and better, information or knowledge the system can discover. This is why we are working on construction of improved opinion representations that permit obtaining additional interesting results, for example, discovering similar opinions, opposite opinions, contradictions, or identifying trends, deviations, or patterns in different opinion components.

References

[Agrawal et al., 1993] Rakesh Agrawal, Tomasz Imielinski and Arun Swami. Database Mining: A Performance Perspective. In *IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning and Discovery in Knowledge-Based Databases*, Vol. 5, No. 6, December 1993, 914-925.

[Agrawal et al., 1996] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant. The Quest Data Mining System. In *Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Portland, Oregon, August, 1996.

[Bhandari et al., 1997] Inderpal Bhandari, Edward Colet, Jennifer Parker, Zacary Pines, Rajiv Pratap, Krishnakumar Ramanujam. Advanced Scout: Data Mining and Knowledge discovery in NBA Data. In *Data Mining and Knowledge Discovery* 1, 121-125, 1997.

[Church and Rau, 1995] Kenneth W. Church and Lisa F. Rau. Commercial Applications of Natural Language Processing. In *Communications of the ACM*, Vol.38, No 11, November 1995.

[Cowie and Lehnert, 1996] Jim Cowie and Wendy Lehnert. Information Extraction. In *Communications of the ACM*, Vol.39, No.1, January 1996.

[Davis, 1989] Roy Davis. The Creation of New Knowledge by Information Retrieval and Classification. In *The Journal of Documentation*, Vol 45, No 4, pp. 273 –301, December 1989.

[Feldman and Dagan, 1995] R. Feldman and I. Dagan. Knowledge Discovery in Textual databases (KDT). In *Proc. Of the 1st International Conference on Knowledge Discovery (KDD_95)*, pp.112-117, Montreal, 1995.

[García-Menier, 1998] Everardo García Menier. Un sistema para la Clasificación de notas periodisticas (in Spanish). In *Proc. Of the Simposium Internacional de Computacion*, CIC-98, México, D. F., 1998.

[Gay and Croft, 1990] Gay, L. and Croft, W. Interpreting Nominal Compounds for Information Retrieval. In *Information Processing and Management* 26(1): 21-38, 1990.

[Glymour et al., 1997] Clark Glymour, David Madigan, Darly Pregibon, Padhraic Smyth. Statistical Themes and Lessons for Data Mining. In *Data Mining and Knowledge Discovery* 1, 11-28, 1997.

[Guzmán, 1998] Adolfo Guzmán. Finding the main Themes in a Spanish Document. In *Expert Systems with Applications*, 14, pages 139-148, 1998.

[Hahn, and Schnattinger, 1997a] Udo Hahn and Klemens Schnattinger. Knowledge Mining from Textual Sources. In F.Golshani & K.Makki (Eds.) CIKM’97, *Proceedings of the 6th International Conference on Information and Knowledge Management.* New York/NY: ACM, Las Vegas, Nevada, USA, November 10-14, 1997, pp.83-90.

[Hahn and Schnattinger, 1997b] Udo Hahn and Klemens Schnattinger. Deep Knowledge Discovery from Natural Language Texts. In D. Heckerman, H.Mannila, D. Pregibon & R.Uthurusamy (Eds.) KDD’97, *Proceedings of the 3rd Conference on Knowledge Discovery* *and Data Mining*. Newport Beach, Cal., August 14-17, 1997. Menlo Park/CA: AAAI Press, 1997, pp.175-178.

[IBM, 1997] IBM. Text Mining: A Quick Overview. *IBM Technology Watch, A Decision Support System*, http: // www. synthema. it / tewat / demo / pres / ntwprese. htm

[Kitani et al., 1994] Tsuyoshi Kitani, Yoshio Eriguchi, and Massami Hara. Pattern Matching and Discourse in Information Extraction from Japanese Text, In *Journal of Artificial Intelligence Research* 2 (1994) 89-100.

[Klavans and Kan, 1998] Judith Klavans and Min-Yen Kan. Role of Verbs in Document Analysis. In Proc. 17th *Conference on Computational Linguistics* (COLING-ACL'98), 1998.

[Lent et al., 1997] Brian Lent, Rakesh Agrawal and Ramakrishnan Srikant. Discovering Trends in Text Databases. In *Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Newport Beach, California, August 1997.

[Schnattinger and Hahn, 1997] Klemens Schnattinger & Udo Hahn. Intelligent Text Analysis for Dynamically Maintaining and Updating Domain Knowledge Bases. In X.Liu, P.Cohen & M.Berthold (Eds.), IDA'97, *Proceedings of the 2nd International Symposium on Intelligent Data Analysis*. London, U.K., August 4-6, 1997. Berlin etc.: Springer, 1997, pp.409-422.

[Weiss and Indurkhya, 1998] Sholom M. Weiss and Nitin Indurkhya. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, Inc., 1998.

[tryb.org site, 1998] http: // www. tryb. org / tmkd / id1_cf.htm, *Text Mining and Knowledge Discovery*, 1998.

Extraction of Document Intentions from Titles

Manuel Montes-y-Gómez,
Alexander F. Gelbukh

mmontesg@susu.inaoep.mx
gelbukh@pollux.cic.ipn.mx
Natural Language Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN).
Av. Juan Dios Bátiz, Zacatenco, 07738 D.F.
Mexico

Aurelio López-López

allopez@inaoep.mx
Electronics, INAOE
Luis Enrique Erro No. 1
Tonatzintla Pue. 72840
México

Abstract

In spite of their small size, titles are a very important source of information about document contents. This is why they are frequently used as a way to obtain document keywords and this is the reason we have chosen to use them to obtain and extract document intentions. In order to construct better document representations, we analyze the opportunities to extract some document details from titles. Particularly, we propose to use some classical information extraction techniques for constructing extratopical representation of the documents. It is put together with the keywords to form a new and more complete representation of the document. A possible use for this representation in the Information Retrieval area is described, as well as how this paradigm for document representation can improve the actual retrieval results.

1 Introduction[*]

Unlike the structured information or formal representations, raw texts have very free and complex form. These characteristics allow them to describe better and more completely all entities and facts, but at the same time these features provoke many analysis difficulties.

Nowadays, almost every raw text operation, for example, text classification, information retrieval, indexing or text description, is done on the basis of keywords or, in the best case, of topics or themes obtained from some parts of the documents or from the entire text [Guzmán, 1998]. This paradigm generally leads to ignoring text characteristics beyond topicality, such as intentions, proposes, plans, content level, etc. [López-López and Myaeng, 1997].

[*] The work of M. Montes-y-Gómez was supported by CONACyT scolarship, México. A. Gelbukh was partially supported by REDII-CONACyT, DEPI-IPN, and SNI, México.

In this paper, we present evidence of the relationship between document intentions and document title. Additionally, we describe a method used for automatic extraction of the document intention(s) and finally proposed a possible use for this information in IR systems.

2 Intention Structure

By *intentions*, we mean determination to do something. Intentions describe, or are related with, the act intended by the document. They are grammatically associated with some verbs that take the main topic of the document as their subject, such as *introduce*, *describe*, *propose*, etc.

The task of determining the document intentions consists in finding verbs whose actions are performed by the document. For instance, we can say that the intention of some document is to *describe* something if there is some evidence in the document that relates the document with the action "describing."

With this approach, extraction of the document intention might seem a simple task; in fact, it is not. Document intentions are more than simple actions related with the documents. They include an action, an object of the action, and sometimes one or more pieces of related information.

For instance, it is not sufficient to say that the intention of some document is to describe, since it is also necessary to indicate what thing the document describes (the object), and it can be necessary to say how, when or why this action is done (some related information).

3 Intentions in Document Titles

Title is not only the very first information the reader receives from a document, but also the part of the document most heavily used for such tasks as indexing and classification. This background inspires us to use titles for extraction of the intentions, We can note the following facts about the relation between titles and intentions [Montes-y-Gómez, 1998]:

- Document intentions are associated with title nominalizations, e.g.: "Numerical solution of the polynomial equation, " "*An Introduction to a Machine-Independent Data Division.*"
- Intentions are also related to some present participle patterns, e.g.: "*Proving theorems by recognition,*" "Computing radiation integrals."

4 Intention Extraction Method

The intention extraction system we developed follows a classic information extraction scheme [Cowie and Lehnert, 1996]. It consists of a tagger, a filtering component, a parser, and a module of generation of output data.

5 Output Representation

The output representation generated by our system is a Conceptual Graph [Sowa, 1983]. This representation is a network of concept nodes and relation nodes, where concept nodes represent entities, attributes, or events, while relation nodes identify the kind of relationship that holds between the concept nodes.

These graphs make it easy to represent the information about the document intentions. This representation permits to easily use this information in many applications. The following graph illustrates the intention of the document (in boldface), and its structure with the additional information, for the first title given above.

$$[\text{numerically}] \xleftarrow{\textit{manner}} \textbf{[solve]} \xrightarrow{\textit{object}} \begin{array}{l}[\text{polinomial-} \\ \text{equation}, \#\,]\end{array}$$

6 Experimental Results

We tested our system on two standard test document collections (CACM-3204 and CISI-1459), consisting of a total of 4663 document descriptions. When comparing the extraction effectiveness against manually identified titles, the method produced a recall and precision of 92%, 98% and 90%, 96%, for CACM and CISI respectively. When analyzing how the intentions from titles complement those identified from abstracts, we achieved as much as 90% of the documents with some kind of intention representation.

7 Conclusions

With this and related works [López-López and Myaeng, 1997; López-López and Tapia-Melchor, 1998; Montes-y-Gómez, 1998], we try to break down the keyword document representation paradigm and begin to use other document characteristics in their representations. In particular, this paper provides evidence for the relations between document titles and their intentions and also demonstrates that these intentions are reflected in titles by some particular nominalizations and present participles patterns. As one of its principal features, the automatic intention extraction method has domain independence, so that it can be applied to documents on any topic.

At this moment we are implementing a new IR system, using a new representation of documents – two-level representation, aimed to improve the information retrieval results, mainly so-called normalized precision. In the future, we plan to apply similar methods to other parts of the documents and develop a better content-detail representation.

References

[Chisnell, *et al*, 1993] Ch. Chisnell, D.V. Rama and P. Srinivasan. Structured Representation of Empirical Information. In *Case-based Reasoning and Information Retrieval,* Technical Report 55-93-07, AAAI Press, 1993.

[Cowie and Lehnert, 1996] Jim Cowie and Wendy Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80-91, January 1996.

[Guzmán, 1998] Adolfo Guzmán Arenas. Finding the Main Themes in a Spanish Document. *Expert Systems with Applications*, 14:139-148, 1998.

[López-López and Myaeng, 1997] A. López-López, and Sung H. Myaeng. Extending the Capabilities of Retrieval Systems by a Two Level Representation of Content. In *Proceedings 1st Australian Document Computing Symposium,* Part I, pages 15-20, Melbourne, Australia, March 1996.

[López-López and Montes-y-Gómez, 1998] A. López-López, and M. Montes-y-Gómez. Nominalization in Titles: A Way to Extract Document Details. In *Proceedings of the Simposium Internacional de Computación CIC'98,* pages 396-404, México D.F. November 1998.

[López-López and Tapia-Melchor, 1998] A. López-López, and Ma. del P. Tapia-Melchor. Automatic Information Extraction from Documents in WWW. In *Proceedings of the VIII International Congress of Electronics, Communications, and Computers CONIELECOMP 98*, pages 287-291, Cholula, Puebla, México. Feb.1998.

[Montes-y-Gómez, 1998] M. Montes-y-Gómez. Information Extraction from Document Titles, M. Sc. Thesis, Electronics, INAOE, México, 1998.

[Sowa, 1983] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley, 1983.

Text Categorization Using a Hierarchical Topic Dictionary

Alexander Gelbukh
Grigori Sidorov
Adolfo Guzmán-Arenas

{gelbukh,sidorov,aguzman}@pollux.cic.ipn.mx
Natural Language Laboratory, Center for Computing Research (CIC),
National Polytechnic Institute (IPN). Av. Juan Días Bátiz, Zacatenco, 07738 DF.
Mexico

Abstract

A statistical method of text categorization driven by a hierarchical topic dictionary is proposed. The method uses a dictionary with a simple structure and is insensible to inaccuracies in the dictionary; the dictionary is easily trainable on a manually classified document collection and even automatically translatable into different languages. A common sense-complaint way of assignment of the weights to the topics is discussed. The discussion is based on the experience with the system CLASSIFIER developed on the base of these methods.

1 Introduction*

We consider the task of text categorization by the topic of the document: for example, some documents are about *animals*, and some about *industry*. In this paper we consider the list of topics to be large but fixed. Our algorithm does not obtain the topics from the document body; instead, it relates the document with one of the topics listed in the system dictionary. The result is, thus, the measure (say, in percents) of the corresponding of the document to each of the available topics.

A problem arises of the optimal, or reasonable, degree of detail for such classification. For example, when classifying the Internet news for an "average" reader, the categories like *animals* or *industry* are quite appropriate, while for classification of articles on zoology such a dictionary would give a trivial answer that all documents are about *animals*. On the other hand, for "average" reader of Internet news it would not be appropriate to classify the documents by the topics such as *mammals*, *herptiles*, *crustaceans*, etc.

* The work done under partial support of DEPI-IPN, CONACyT (26424-A), REDII-CONACyT, and SNI, Mexico.

In this paper, we will discuss the structure of the topic dictionary, the choice and use of the weights of individual nodes in the hierarchy, and some practical aspects of compilation of the topic dictionary.

2 Topic hierarchy and classification algorithm

In [Guzmán-Arenas, 1997; 1998] it was proposed to use a hierarchical dictionary for determining the main themes of a document. Technically, the dictionary consists of two parts: *keyword groups* representing individual topics, and a *hierarchy* of such topics.

A keyword group is a list of words or expressions related to the situation referred to by the name of the topic. For example, the topic *religion* lists the words like *church*, *priest*, *candle*, *Bible*, *pray*, *pilgrim*, etc. Note that these words are connected neither with the headword *religion* nor with each other by any "standard" semantic relation, such as subtype, part, actant, etc.

The topic tree organizes the topics, as integral units, into a hierarchy or, more generally, a lattice (since some topics can belong to several nodes of the hierarchy).

The algorithm of application of the dictionary to the task of topic detection also consists of two parts: individual (leaf) topic detection and propagation of the topic weights up the tree.

The first part of the algorithm is responsible for detection individual (leaf) topics. Effectively, it answers, topic by topic, the following question: To what degree this document corresponds to the given topic? Such a question is answered for each topic individually. For more information on how the topic weights are determined (in a slightly different situation), see [Alexandrov and Gelbukh, 1999]. In the simplest case, the weight of a topic is the number (frequency) of words from the corresponding word list, found in the document.

The second part of the algorithm is responsible for propagation of the found frequencies up the tree. With this, we can determine that a document mentioning the leaf topics *mammals*, *herptiles*, *crustaceans*, is relevant for the non-leaf topic *animals*, and also *living things* and *nature*.

The question discussed in the paper is how far are to be propagated the weights up to the tree, for the determined main topic of the document not to be trivially general, like *objects*.

3 Relevance and discrimination weights

Instead of simple word lists, some numeric weights can be used by the algorithm to define (1) the quantitative measures of relevance of the words for topics and (2) the measure of importance of the nodes of the hierarchy.

The first type of weights, which we call *relevance weights*, is associated with the links between words and topics and the links between the nodes in the tree. For example, if the document mentions the word *carburetor*, is it about *cars*? And the word *wheel*? I.e., how relevant is the word *carburetor* or *wheel* for the topic *cars*, how strong is their relationship? Intuitively, the contribution of the word *carburetor* into the topic *cars* is greater than that of the word *wheel*; thus, the link between *wheel* and *cars* is assigned a less weight.

It can be shown that the weight w_k^j of such a link (between a word *k* and a topic *j*, or between a topic *k* and its parent topic *j* in the tree) can defined as the mean relevance for the given topic of the documents containing this word: $w_k^j = \sum_{i \in D} r_i^{\,j} n_i^k \Big/ \sum_{i \in D} n_i^k$. Here the summation is done by all the available documents *D*, $r_i^{\,j}$ is the measure of relevance of the document *i* to the topic *j*, and n_i^k is the number of occurrences of the word or topic *k* in the document *i*.

Unfortunately, we are not aware of any reliable algorithm to automatically detect the measure of relevance $r_i^{\,j}$ of the documents for the domains in an independent way. For the moment, such a measure is estimated manually by the expert, and then the system is trained on the set of documents. Alternatively, the expert can usually intuitively assign the relevance weights to the documents.

Both these approaches require manual work. To avoid it, as a practical approximation, for narrow enough themes the hypothesis can be assumed that the texts on this topic almost never occur in general texts (newspaper mixture). Then the expression for the weights can be simplified: $w_k^j = 1 \Big/ \sum_{i \in D} n_i^k$.

The main requirement for the second type of weights – the *discrimanation weights* – is their *discrimination power*: a topic should correspond to a (considerable) *subset* of documents. On the other hand, the topics that correspond to nearly all the documents in the data base are useless because they do not permit to make any relevan conclusions about the corresponding documents.

Thus, the weight w^j of a tree node *j* can be estimated as the variation of the relevance $r_i^{\,j}$ the topic over the documents of the database. A simple way to calculate such a discrimination power is to measure it as the dispersion: $w^j = \sum_{i \in D} \left(r_i^{\,j} - M\right)^2$, where *M* = $\sum_{i \in D} r_i^{\,j} \Big/ |D|$ is the average value of $r_i^{\,j}$ over the current database *D*, and $r_i^{\,j}$ is determined by the former algorithm, i.e., without taking into account the value of w^j . In a more precise manner, the information theory can be applied to the calculation of the weights; we will not discuss here this idea.

With this approach, for, say, a biological database, the weight of the topics like *animals*, *living things, nature* is low because all the documents equally mention these topics. On the other hand, for the newspaper mixture their weight is high, since many documents in it do not correspond to these topics, but still some considerable part do.

References

[Alexandrov and Gelbukh, 1999] M. Alexandrov and A. Gelbukh. Measures for Determining Thematic Structure of Documents with Domain Dictionaries. In *Proc. Workshop on Text Mining, International Joint Conference on Artificial Intelligence IJCAI-99,* to appear, 1999.

[Guzmán-Arenas, 1997] Adolfo Guzmán-Arenas. Hallando los temas principales en un artículo en español (in Spanish). *Soluciones Avanzadas,* 5(45):58, 5(49):66, 1997.

[Guzmán-Arenas, 1998] Adolfo Guzmán-Arenas. Finding the main themes in a Spanish document. Journal Expert Systems with Applications, 4(1/2):139-148,. January-February 1998.

[Gelbukh et al., 1997] Alexander Gelbukh, Igor Bolshakov, Sofía Galicia Haro. Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts. In *Proceedings of the International Conference on Automatic Learning and Discovery*, Pittsburgh, PA, USA, June 1998. Carnegie Mellon University.