Text Mining as a Social Thermometer

Manuel Montes y Gómez [1]
Aurelio López López [2]
Alexander F. Gelbukh [1]

[1] CIC, IPN,
Laboratorio de Lenguaje Natural.
Av. Juán de Dios Bátiz, México DF.
Tel. +52 (5) 729-60-00, ext. 56544.
e-mail: mmontesg@susu.inaoep.mx
gelbukh@pollux.cic.ipn.mx

[2] INAOE, Electronics.
Luis Enrique Erro No. 1
Tonantzintla, Puebla, 72840 México.
Tel. (52 22) 472-011 Fax (52 22) 470-517
e-mail: allopez@gisc1.inaoep.mx

Abstract

"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" [http://www.tryb.org/tmkd/id1_cf.htm].

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 distributions analysis, average calculus and standard deviation computation to discover information about 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 of time.

1 Introduction

Without a doubt, newspapers and Internet news remain one of the most important information media that reflects 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 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 & Dagan, 1995; Lent, Agrawal & Srikant, 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 [Felman & Dagan, 1995].

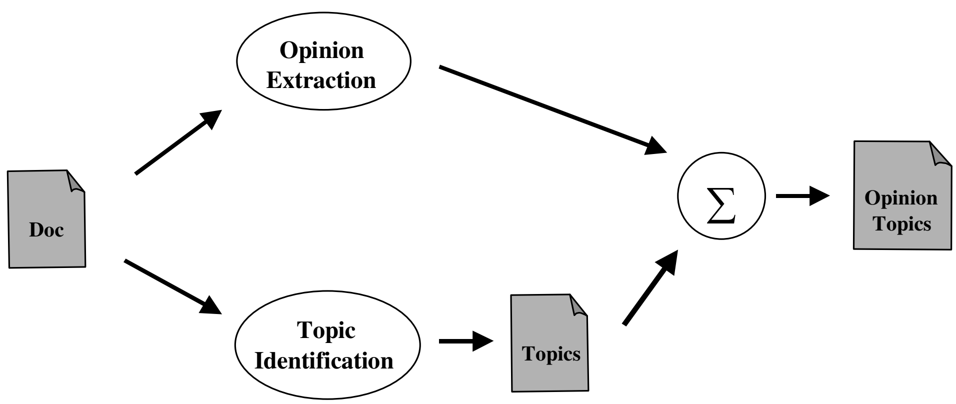


Figure 1. Opinion topic extraction system.

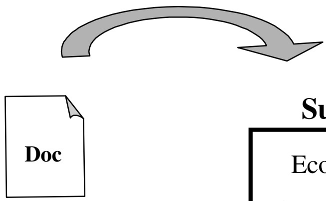
Considering this problem and the way others have resolved it, we designed an opinion acquisition method based on well-known indexing and information extraction techniques. Figure 1 shows the architecture of our opinion topic extraction system.

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 & 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, Eriguchi & Hara, 1994] using as a trigger a list of verbs denoting communication actions, such as Spanish *dijo* 'say', *propuso* 'propose', etc. [Klavans & Yen Kan, 1998].

The third module matches topics with opinion paragraphs and selects only the topics that are explicitly mentioned in this 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, i.e. *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.



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.

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:

1. The trend discovery.
2. The 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 the change trend case, it is important to discover the main change sources, for instance, the opinion topics with the maximum change rates. In the 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, Madigan, Pregibon & Smyth, 1997]. These distributions have been used before for the same purpose [Feldman & 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 the 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 and 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 and f_i^2 be the frequencies of the opinion topics (θ_i) at the times t_1 and t_2.

<u>*Integration*</u>:

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

$$T = T_1 \cup T_2$$

Filtering:

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

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

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

$$f'^k_i = \begin{cases} f_i^k & \text{if } \theta_i \in T_k \\ 0 & \text{otherwise} \end{cases}$$

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

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. The 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.

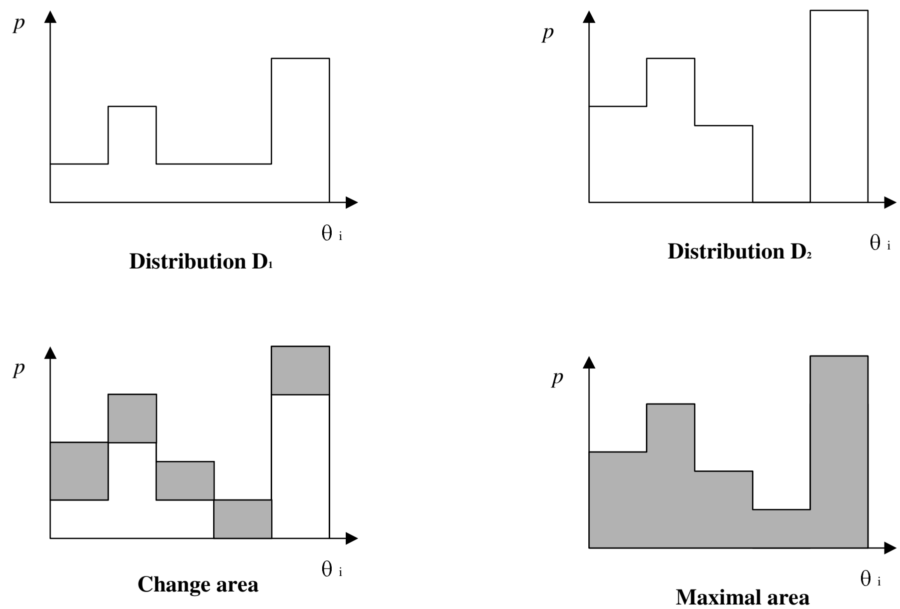


Figure 3. Comparison method.

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.

$$\forall \theta_i \in T':$$

$$p_i^k = \frac{f'^k_i}{\sum_j f'^k_j} \qquad k \in \{1,2\} \quad \textit{Frequency value of topic } \theta_i$$

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

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

$$dF_\sigma = \sqrt{\frac{\sum (dF_i - dF_\mu)^2}{N}} \qquad \textit{Difference desviation}$$

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

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

This criterion was found empirically, but in other fields (or maybe 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:

$$T'' = \{\theta_i \in T' \mid (f'^{1}_{i} >> f'^{1}_{\mu}) \; and \; (f'^{2}_{i} >> f'^{2}_{\mu})\}$$

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

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

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:

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

$$\text{where } T' = \{\theta_i \mid f'^{1}_{i} + f'^{2}_{i} > \beta = 1\}$$

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

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

Topics	***f´ 1***	***f´ 2***	***p´ 1***	***p´ 2***	***dF***
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

Trend Discovery:

$$A_c = 1.017$$
$$A_m = 1.566$$
$$\mathbf{C_r} = \mathbf{0.65}$$

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

5 Conclusions and Future Work

These experiments and results encourage us to continue working in this direction. We showed 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 a further work, we plan to:

1. 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.

2. 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?
3. Analyze and classify the opinions on types. For example, opinions in which something is proposed, opinions is prognosticated, qualified, etc. This classification could be interesting and useful for a high level analysis of opinions.

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

Acknowledgments

This investigation was partially funded by a scholarship granted by Consejo Nacional de Ciencia y Tecnología, Centro de Investigación en Computación (CIC-IPN), and REDII-CONACyT, Mexico.

References

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

[Guzman, 1996] Adolfo Guzmán, Uso y diseno de Mineros de Datos, Soluciones Avanzadas, num. 34, 1996.

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

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

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

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

[Freund & Walpole, 1990] Freund y Walpole, Estadística Matematica con Aplicaciones, Cuarta Edición, Prentice Hall, 1990.

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

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

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

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

[Schnattinger & 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.

[Agrawal, Imielinski & Swami, 1993] Rakesh Agrawal, Tomasz Imielinski and Arun Swami, Database Mining: A Performance Perspective. 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, Arning, Bollinger, Mehta, Shafer, Srikant, 1996] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant, The Quest Data Mining System, *Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Portland, Oregon, August, 1996.

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

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

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

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

[Hahn, & Schnattinger, 1997] Udo Hahn, & 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 & Schnattinger, 1997] Udo Hahn & 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.

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