Mining the News: Trends, Associations, and Deviations[*]

Manuel Montes-y-Gómez[1], **Alexander Gelbukh**[1], and **Aurelio López-López**[2]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, esq. Mendizabal,
Zacatenco, 07738, México D.F., Mexico.
e-mail: mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

[2] INAOE
Luis Enrique Erro No. 1
Tonantzintla, Puebla, 72840 Mexico.
Tel. (52 22) 472011 Fax (52 22) 470517
e-mail: allopez@inaoep.mx

Article received on August 28, 2000; accepted on February 16, 2001

Abstract

News reports are an important source of information about society. Their analysis allows to understand its current interests and to measure the social importance of many events.

In this paper, we use the analysis of news as a means to explore the society interests. We present a text mining technique that uncovers trends, discovers associations and detects deviations from news notes. The method uses simple statistical representations of the news reports (frequencies and probability distributions of topics) and statistical measures (the average or median, the standard deviation, and the correlation coefficient) for analysis and discovery of useful information.

We illustrate the method with some results obtained from preliminary experiments and discuss their main implications.

Keywords : text mining, data mining, society interests, text processing, news analysis, statistics.

1 Introduction

The problem of analysis of large amounts of information has been solved to a good degree for the case of information that has a fixed structure, such as databases. The methods of statistical analysis of large databases are called data mining (Fayyad *et al*., 1996; Han and Kamber, 2001). However, this problem remains unsolved for non-structured information such as unrestricted natural language texts.

Text mining has emerged as a new area of text processing that attempts to fill this gap (Feldman, 1999; Mladenic, 2000). It can be defined as data mining applied to textual data, i.e., as discovery of new facts and world knowledge from large collections of texts that – unlike the problem of natural language understanding – do not explicitly contain the knowledge to be discovered (Hearst, 1999). Naturally, the goals of text mining are similar to those of data mining: for instance, it also attempts to uncover trends, discover associations and detect deviations in a large set of texts. Text mining has also adopted techniques and methods of data mining, e.g., statistical techniques and machine learning approaches.

We present a method for analyzing the collections of news appearing in newspapers, newswires, or other mass media. Analysis of news collections is an interesting challenge since news reports have many characteristics different from the texts in other domains. For instance, the news topics have a high correlation with society interests and behavior, they are very diverse and constantly changing, and also interact and influence each other.

[*] Work done under partial support of CONACyT, REDII-CONACyT, and SNI, Mexico.

The method we describe is adapted for these characteristics of our domain. It uses simple statistical representations for the news reports (frequencies and probability distributions) and statistical measures (the average or median, the standard deviation, and the correlation coefficient) for the analysis and discovery of useful information (Glymour *et al*., 1997): trend analysis, deviation detection, and discovery of ephemeral associations between news topics.

The process of text mining is usually divided into two stages (Ah-Hwee Tan, 1999):

1. Preprocessing the documents, that is, transforming them into a structured formal representation, for instance, a list of keywords or topics;
2. Discovering interesting facts and new knowledge, such as trends, associations, and deviations, from the resulting structured data collection.

The structured representations and thus the methods used in their processing vary significantly from application to application. In case of deep analysis of text, the complex structure of the data being analyzed forces the researchers to simultaneously develop new methods specific for the text mining tasks (Montes-y-Gómez *et al*., 2000). However, purely topical analysis done at the word level allows a better reuse of the methods developed in frame of data mining (Feldman & Dagan, 1995; Feldman & Hirsh, 1996; Feldman *et al*., 1998**;** Rajman & Besançon, 1998; Shian-Hua Lin *et al*., 1998; Lent *et al*., 1997; Nahm and Mooney, 2000).

In this paper, we follow the latter idea. As the structured representation of the contents of a news report, we use a list of keywords or topics with their respective frequencies (numbers of occurrences). Once the news reports are represented with such a list of topics, we analyze them and attempt to discover some interesting facts, mainly trends, associations, and deviations, that contribute to a better understanding of the society behavior and interests.

Trend analysis answers the questions like: What is the general trend of the society interests (news topics) between the two periods? Is there a significant change in the news topics? Are the news topics almost the same in these two periods? What are the emerging or disappearing topics? What topics did not change their importance? What topics have a behavior significantly contrary to the general trend?

Ephemeral association discovery focuses on the analysis of a very common phenomenon in news, that is, the influence of the peak news topics on other topics. Here we try to answer the questions like: Which topics emerged along with the peak topic? Which topics were temporarily forgotten when the peak topic appeared?

Deviation detection focuses on irregularities, mainly on detecting news reports that differ from the typical case in their topics, as well as on detecting the specific sources of news flows – say, newspapers or newswires – that differ in their topic profiles from the majority of other such sources. Such deviations can have interesting social implications. Mainly, we are interested in answering the questions like: Which newspaper did not mention on the front page a topic that was a front page news in other newspapers at the same place (city, country) in the same time? Which newspaper did include a subject that the other newspapers did not mention? Which newspapers frequently exhibit such special behavior?

In the rest of the paper, after introduction of the necessary mathematical notions, we explain the methods we suggest for trend, association, and deviation analysis of the described types. Then, we illustrate each of these methods with real-world examples, and finally formulate our conclusions.

2 Construction of Probability Distributions

Given a collection of news reports corresponding to some time span of interest, we construct a structured representation of each news report. This representation consists of the information on its source, date, author, etc., and a formal representation of its contents.

For the latter, we reduce the text to a list of keywords, or topics. In our experiments, we used a method similar to one proposed by Gay and Croft (1990), where the topics are related to noun strings. We apply a set of heuristic rules specific for Spanish and based on proximity of words that allow identifying and extracting phrases. These rules are driven by the occurrence of articles and the preposition *de* (‘of’) along with nouns or proper names; some morphological inflection patterns (typical endings of nouns and verbs) are also taken into account. For instance, given the following paragraph, the algorithm selects the highlighted words as keywords:

“***Góngora Pimentel*** aseguró que estas ***demandas*** se resolverán en un ***plazo*** no mayor de 30 días y que sin duda la ***demanda interpuesta*** por el ***PRD*** se anexará a la que presentó el ***Partido Acción Nacional***”.[1]

Once this is done, a frequency f_k^i is assigned to each topic discussed in the period of interest. It is calculated as the number of the news reports in the period i that mention the topic k.

Then, using these frequencies of the topics, a probability distribution $D_i = \left\{ p_k^i \right\}$ of the news topics in the period i

[1] ‘***Góngora Pimentel*** confirmed that these ***demands*** will be satisfied in a ***period*** not longer than 30 days and that without any doubt the ***demand introduced*** by ***PRD*** will be added to that presented by the *National Action Party*.’

is constructed, where p_k^i expresses the probability of occurrence of the topic k in the period i:

$$p_k^i = \frac{f_k^i}{\sum_{j=1}^{n} f_j^i}$$

Here, n is the number of topics cited in the whole period i.

In practice, each distribution D_i is built as a sparse vector, i.e., as a list of pairs ($topic_k$, p_k^i) with only the topics actually mentioned in that period (i.e., $p_k^i \neq 0$) being physically included in the list, see Table 1 below. For technical reasons, before any calculations, two operations are performed on these lists: filtering and merging.

Filtering. Such lists are usually very large. To focus our analysis in the main society interests, the topics not significant for the analysis – those describing the noise of one-time events and not clearly related to the main interests of society – are removed from the lists, so that the lists include only the topics such that $f_k^i > b$ for some threshold b that specifies the minimum frequency for a topic to be considered interesting; its value is determined empirically.

Merging. Before any comparison, the lists are merged to describe the same topics (to have the same length). This is achieved by adding (with zero values) to each list the topics present in the other list: $D'_i = \{(topic_k,\ p_k^i) \mid p_k^1 \neq 0$ or $p_k^2 \neq 0\}$. Now that the new lists D'_1 and D'_2 have the same length, these lists technically can be operated upon in the formulas like (1) below as usual vectors rather than sparse vectors, ignoring the first member $topic_k$ in each pair.

Once this operation is done, the topic probabilities are re-normalized so that the new values of p_k^i give $\sum_{i=1}^{n} p_k^1 = \sum_{i=1}^{n} p_k^2 = 1$.

3 Comparison and Trend Analysis

The main goal of trend analysis is to study the behavior of the society interests, i.e., determining if they change or remain considerably stable from one period to another. Since the behavior of only one of them is not always a good sign of their general behavior, we use a method that considers all society interest at a certain time, in order to detect their general trend. The method we use is a general method for comparing two news collections: two news sources, the newspapers from two geographic places, etc. We apply it to comparing the news collections covering two different periods of time, which gives us their temporal trend.

Following this idea, we divide the trend analysis task in two ones: general trend discovery and identification of the topics (factors) that contribute to this trend. Trend discovery attempts to determine if the society interests are noticeably different or similar between two time spans.[2] Since we represent society interests for a given period as the set of all news topics mentioned in it, we uncover a trend in the society interests by determining if the sets of news topics present a considerable change or almost remain stable between the two periods.

Once a trend is uncovered, it is important to determine which topics contributed most significantly to the trend. In case of a change trend, it is important to discover the main sources of the change, for instance, the topics with the highest change rates. In case of stability trends, it is important to identify the stability factors, for instance, those of the most actively discussed topics that remained without significant change.

Additionally, we detect some topic deviations, that is, some topics having a behavior significantly contrary to the general trend.

3.1 Comparison. Trend discovery

We discover trends by comparing the probability distributions $D_i = \{ p_k^i \}$ of the news topics for two given periods[2] $i = 1, 2$. Probability distributions had been used for the same purpose (Feldman & Dagan, 1995; Lent *et al*., 1997) but with a different similarity measure, e.g. Feldman & Dagan (1995) used the relative entropy measure (KL-distance).

Since the KL-distance is not symmetric while we are interested in the change regardless of the direction and a reference information source, we use a different comparison measure C_c for two distributions: the quotient of the change area and the maximal area, see Figure 1. This measure reflects an overall trend since we focused on general society interests. It does not measures individual proportions of change of each individual factor.

$$\begin{array}{ll} C_c = \dfrac{A_c}{A_m} & \text{change coefficient, where:} \\ A_c = \displaystyle\sum_{k=1}^{n} d_k & \text{change area} \\ A_m = \displaystyle\sum_{k=1}^{n} \max\left(p_k^1,\ p_k^2\right) & \text{maximal area} \\ d_k = \left|p_k^1 - p_k^2\right| & \text{individual topic change} \end{array} \tag{1}$$

If the change coefficient between two probability distributions is greater than some user-specified threshold γ, i.e., $C_c > \gamma$, then there exists a global change trend between the two given periods; if $C_c < \gamma$, then there exists a stability trend, i.e., there is no significant differences between the two periods. The typical numerical values for user-defined parameters such as γ are given in Section 6.

[2] Or to two sources, geographic locations, etc.

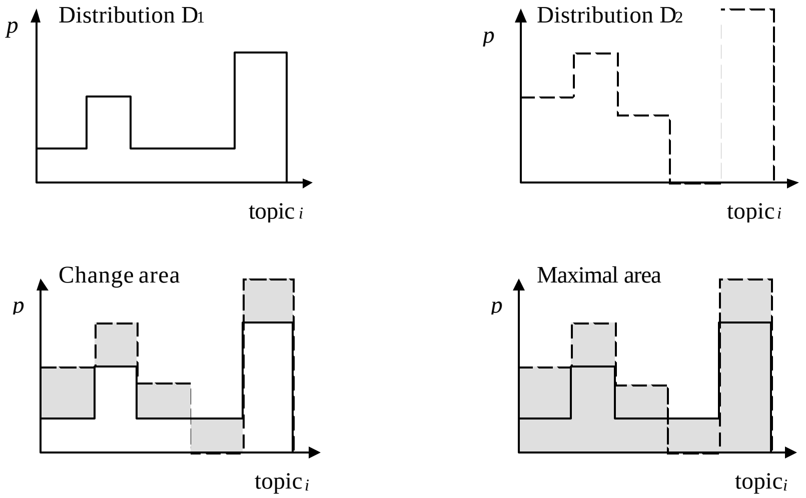


Figure 1. The comparison method

3.2 Identification of change factors

A global change trend is caused essentially by abrupt changes d_k of some individual topics; see (1) above. These changing topics are what we call change factors. We define them as the topics with a change noticeably greater than the typical change. Let d_m be a "typical" value of d_k (see below) and d_s be a measure of the "width" of the distribution. Then the topics $topic_k$ for which $d_k > d_m + (C \times d_s)$ are identified as a change factors. The tuning constant C determines the criterion used to identify an individual change as noticeable.

There are different ways to define the "typical" value d_m and the "width" measure d_s. The most straightforward way is to define them as the average and standard deviation, correspondingly:

$$d_\mu = \frac{1}{n}\sum_{k=1}^{n} d_k \qquad \text{average change}$$

$$d_s = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(d_k - d_\mu\right)^2} \qquad \text{standard deviation of the change}$$

Another way is to define them as a median and the first moment $\boldsymbol{E}[\, d_k - d_\mu \,]$ relative to the median, correspondingly (Cramér, 1999). If:

$$x_1 = \min_x \left\{ x \text{ such that} \sum_{d_k > x} d_k \le \frac{1}{2}\sum_{k=1}^{n} d_k \right\}$$

$$x_2 = \max_x \left\{ x \text{ such that} \sum_{d_k > x} d_k \ge \frac{1}{2}\sum_{k=1}^{n} d_k \right\}$$

then median and the first moment $\boldsymbol{E}[\, d_k - d_\mu \,]$ relative to the median are defined as:

$$d_\mu = \frac{x_1 + x_2}{2} \qquad \text{median change}$$

$$d_s = \boldsymbol{E}\left[d_k - d_\mu\right] = \frac{1}{n}\sum_{k=1}^{n}\left|d_k - d_\mu\right| \qquad \text{moment relative to median}$$

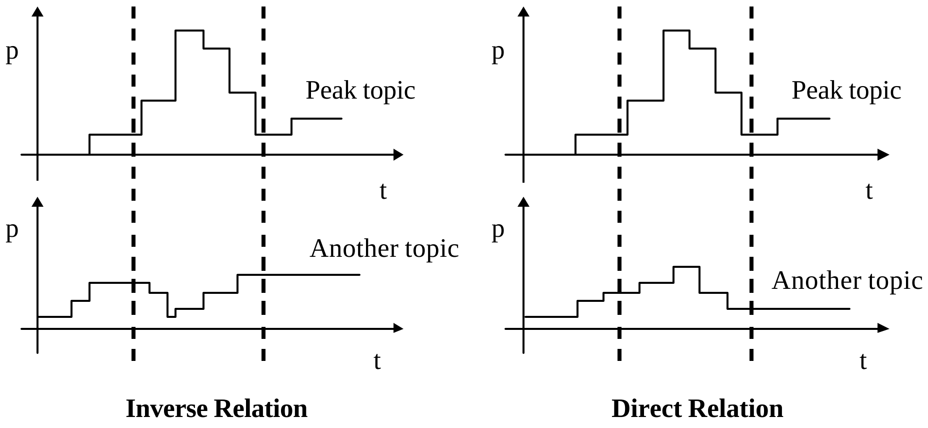
The median can be easily computed by ordering the values and then summing them starting from the greatest value until half the total sum is reached.

The first method – average and standard deviation – has a clear mathematical meaning[3] and is more intuitive in the sense that the en users can easier interpret the results and easier adjust the parameter C. On the other hand, this method is too sensible to the threshold b used for filtering the data (see Section 2). If the value of the threshold is low, the great number of topics with near-zero probabilities will greatly affect the results.

With such very unbalanced distributions that have a great number of near-zero elements, median works better than average, since it is not affected by zero probability elements, however great their number. On the other hand, it is more expensive computationally and less intuitive for the end users.

Choosing between these two methods needs further investigation and probably depends on the specific task.

[3] The Chebyshev theorem states that at least $1 - 1/k^2$ percentage of the data falls into k standard deviations from the average, with $k > 1$. Experimentally, we have determined that a good value for C to be 1. Possibly this is because the distributions in our experiments were small and quasi-homogeneous.



Inverse Relation

Direct Relation

Figure 2. Ephemeral associations between news topics

3.3 Identification of stability factors

Generally speaking, a stability trend is produced by all topics, but the most discussed topics are those contributing more significantly to produce this trend. As the important stability factors, we select the topics that remain almost stable and maintain significant level of importance in both periods. Thus, a *topic*$_k$ is a stability factor if $d_k < d_m - (C \times d_s)$ and $p_k^i > p_m^i$ for both periods *i* = 1, 2. Here, the constant *C* determines the criterion to identify an individual topic as sufficiently stable (a typical value of *C* is 1) and p_m^i is the average probability of the topics in the period *i* defined as:

$$p_m^i = \frac{\sum_{j=1}^{n} p_j^i}{n}$$

where *n* is the number of topics mentioned in the period *i*.

Again, here for p_k^i, the median can be used since it is not affected by the noise of the numerous topics with very low frequency. As we have mentioned above, the issue of choosing one of these variants needs further investigation.

3.4 Topic deviations

The trend of society interests indicates the general behavior of the news topics for the two periods. A topic having a behavior significantly contrary to this general behavior is a news topic deviation.

Then, when there is a stability trend in the society interests, a news topic *k* can be considered a deviation if: (1) its change was noticeably greater than the typical (average or median) change, i.e., $d_k > d_m + (C \times d_s)$ and (2) there would exist a change trend if all topics had the change magnitude *d*$_k$, i.e., $(nd_k / A_m) > g$.

When there is a change trend, a news topic *k* can be considered a deviation if: (1) it remains almost constant in the two periods, that is, its change is considerably less than the typical change (i.e., $d_k < d_m - (C \times d_s)$) and (2) there would exist a stability trend if all topics had the change magnitude *d*$_k$, i.e., $(nd_k / A_m) < g$.

In addition, when there is more than one topic deviation, we mark as the most important those ones that have significant level of importance in both periods, that is, those satisfying the condition $p_k^i > p_m^i$ for both periods *i* = 1, 2.

4 Ephemeral Association Discovery

A usual phenomenon in news is the influence of a peak news topic, i.e., a topic with one-time short-term peak of frequency, over the other news topics. This influence shows itself in two different forms: the peak topic induces some topics to emerge or become important along with it, and the others to be momentarily forgotten.

This kind of influences (time relations) is what we call ephemeral associations.[4] An ephemeral association can be viewed as a direct or inverse relation between the probability distributions of the given topics over a fixed time span. Figure 2 illustrates these ideas and shows an inverse and a direct ephemeral association occurring between two news topics. A direct ephemeral association indicates that the peak topic probably caused the momentary arising of the

[4] This kind of associations is different from the associations of the form $X \Rightarrow Y$, because they not only indicate the coexistence or concurrence of two topics or a set of topics (Ahonen-Myka, 1999; Rajman & Besançon, 1998; Feldman & Hirsh, 1996), but mainly indicate how these news topics are related over a fixed time span.

other topic, while an inverse ephemeral association suggests that the peak topic probably produces the momentary oblivion of the other news topic.

The statistical method we use to detect this ephemeral association is the correlation measure r (Freund and Walpole, 1990), expressed as:

$$r = \frac{S_{01}}{\sqrt{S_{00}S_{11}}},$$

$$S_{kl} = \sum_{i=1}^{n}\left(p_k^i p_l^i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} p_k^i\right)\left(\sum_{i=1}^{n} p_l^i\right),$$

$$k, l = 0, 1.$$

Here p_0^i is the probability of the peak topic and p_1^i of the other topic in the period i.

The correlation coefficient r measures how well two variables are related to each other.[5] It takes values between -1 and 1, where -1 indicates that there exists an exact inverse relation between the two variables (news topics in this case); 1 indicates the existence of an exact direct relation between the variables and 0 the absence of any relation at all.

Thus, if the correlation coefficient between the peak topic and other news topic is greater than a user-specified threshold u, that is, if $r > u$, then there exists an observable *direct* ephemeral association between the topics. Moreover, if the correlation coefficient is less than the threshold $-u$, that is, if $r < -u$, then there exists an *inverse* ephemeral association between the topics.

There are two reasons for introducing the user-specified threshold u: first, to soften the criterion so that we can approximate the way a human visually detects the association, and second, to take care of the relatively small data sets in our application: since we have few data, one or two outliers greatly affect the value of the correlation coefficient.

There is an interesting issue in the interpretation of inverse ephemeral relations between the topics: they can indicate either real-world changes in the corresponding activities (when a war begins, there is less activity in football) or just the effect of limited capacity of the news media (during the election campaign, the newspapers just do not have place on their front pages for football-related reports). See the corresponding discussion in (Montes-y-Gómez *et al*., 2001), where the difference between the observable (as is described here) and the real-world associations is discussed and incorporated in the method.

5 Deviation Detection

The detection of deviations in huge collections of data is an important, but difficult task. It aims at discovering irregular elements in a great amount of data.

In data mining and text mining, detection of deviations is defined as the discovery of something out of the norm, i.e., the detection of anomalous instances that do not fit into the standard case or cases (Knorr *et al*., 2000; Arning *et al*., 1996; Feldman & Dagan, 1995). In many cases, this norm has been a representation of the average element. For instance, Feldman and Dagan (Feldman & Dagan, 1995) consider a topic to be a deviation if this topic has a probability distribution significantly different from the average probability distributions of its siblings (similar topics, i.e., leafs of the same node in a hierarchy).

Following this strategy, we have designed a method for detecting irregular news sources (collection of news, e.g., newspapers). We define a deviation source as one that differs in their topic profile from the other news sources. In particular, we determine that a newspaper is a deviation if it has a noticeable difference in *one* topic with respect to the average of the newspapers in hand. That is, if we have a set of newspapers $N = \{n_i\}$ and represent each one of them as a probability distribution,[6] $D_{n_i} = \{p_k^{n_i}\}$, where $p_k^{n_i}$ is the probability of occurrence of the topic k in the newspaper n_i, p_k^m express the average probability for the topic k in all newspapers, and p_k^s the standard deviation for the same topic, then the newspaper n_x is a deviation if for some topics k,

$$\left| p_k^{n_x} - p_k^m \right| > C \times p_k^s$$

Here again, the user-specified constant C determines the criterion to identify an individual probability as noticeably different from the others.

Another criterion for detecting deviations consists in detecting the elements in the collection unique with respect to some property, which in our case is the fact that they do or do not mention some topic. This criterion tends to give more restricted results, and the deviations found tend to be more interesting.

According to this criterion, a newspaper is a deviation if it differs from the rest of the newspapers in one of the topics. For instance, a newspaper is a deviation if it, say, mentions at the front page a topic that the other newspapers do not, or if it does not mention something that is mentioned by all the rest of the newspapers.

[5] The usual interpretation of the correlation coefficient is that $100r^2$ is the percentage of the variation in the values of one of the variables that can be explained by the relation with the other variable.

[6] These distributions represent the same set of news topics. Therefore, a probability $p_k^{n_i} = 0$ indicates that the topic k was not discussed by the newspaper n_i.

k	**Topics**	f_k^1	f_k^2	p_k^1	p_k^2	d_k
1	*Bancos* (Banks)	7	4	0.212	0.125	0.087
2	*Meta inflacionaria* (inflationary goal)	3	0	0.090	0	0.090
3	*Política monetaria* (Monetary policy)	4	4	0.121	0.125	*0.004*
4	*Ajuste Fiscal* (Fiscal adjustment)	2	0	0.060	0	0.060
5	*Inflación* (Inflation)	4	0	0.121	0	***0.121***
6	*Unión monetaria* (Monetary union)	2	0	0.060	0	0.060
7	*Tasa de interés* (Interest rate)	3	9	0.090	0.280	***0.190***
8	*Política fiscal* (Fiscal policy)	2	0	0.060	0	0.060
9	*Economías asiáticas* (Asian economies)	1	1	0.030	0.031	*0.001*
10	*Brasil* (Brazil)	1	4	0.030	0.125	0.095
11	*Economía nacional* (National economy)	2	1	0.060	0.031	0.029
12	*Cambio de moneda* (Change of currency)	0	3	0	0.094	0.094
13	*Bolsa de valores* (Stock market)	2	2	0.060	0.062	*0.002*
14	*Crisis financiera* (Financial crisis)	0	2	0	0.062	0.062
15	*Mercado financiero* (Financial market)	0	2	0	0.062	0.062

Table 1. Data for the trend analysis

Thus, given a set of newspapers $N = \{n_i\}$ and representing each one of them as a list of keywords or topics, $n_i = \{topic_k\}$, we call the newspaper n_x a deviation if any one of the following holds:

1. It mentions a topic that none of the other newspapers does,

$$\exists\, topic_y : (topic_y \in n_x) \wedge \left(topic_y \notin \bigcup_{i \neq x} n_i \right)$$

2. It does not mention a topic that all other newspapers do.

$$\exists\, topic_y : (topic_y \notin n_x) \wedge \left(topic_y \in \bigcap_{i \neq x} n_i \right)$$

6 Experimental results

6.1 Trend analysis

As an example, let us consider the economic news from *El Universal*,[7] a Mexican newspaper, for the last week of January and for the first week of February of 1999. There are 47 different topics in these two weeks, but after merging and filtering (see Section 2) we get only 15 topics (b = 1 was used). Table 1 shows these topics and their probabilities.

For this collection, A_c = 1.017, A_m = 1.504, and C_c = 0.676. Using $g = 0.5$, we conclude that there is a global change trend between these periods.

Since d_m = 0.0678 and d_s = 0.048 for the given topic set, using C = 1, the change factors discovered are: *inflation* as a disappearing topic, and *interest rate* as an emerging topic.

We can detect that the news topics *monetary policy*, *Asian economies*, and *stock market* are deviations, the most important deviation being the topic *monetary policy*. This is because it almost remained constant, d_k = 0.004, while at the same time it was an important topic in both periods (its probabilities were greater than the average probabilities in both periods).

6.2 Ephemeral association discovery

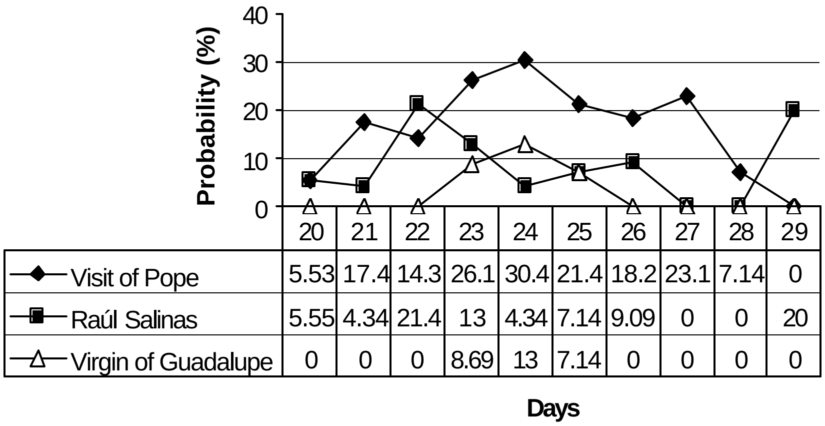
As another example, let us consider the national news from the Mexican newspaper *El Universal* for the ten days surrounding the visit of Pope John Paul II to Mexico City, i.e. from January 20 to 29 of 1999. The topic *visit of Pope* is a peak topic in this period.

We detect two interesting ephemeral associations. A direct association holds between the peak topic and the topic *Virgin of Guadalupe*,[8] with a correlation coefficient r = 0.959 for the period between January 23 and January 25 (stay of the Pope in Mexico), and r = 0.719 for the period between the 20 and 29 of January. Also, there is an inverse association between the peak topic and the topic *Raúl Salinas* (brother of the Mexican ex-president, Carlos Salinas de Gortari, sentenced in those days), with a correlation coefficient r = – 0.703 between the 22 and 26 of January (period covering the visit of Pope and the sentencing of Raúl Salinas).

The direct association between the peak topic and the topic *Virgin of Guadalupe* indicates that probably the topic

[7] http://www.el-universal.com.mx

[8] A Mexican saint whose temple the Pope visited.



	20	21	22	23	24	25	26	27	28	29
◆ Visit of Pope	5.53	17.4	14.3	26.1	30.4	21.4	18.2	23.1	7.14	0
■ Raúl Salinas	5.55	4.34	21.4	13	4.34	7.14	9.09	0	0	20
△ Virgin of Guadalupe	0	0	0	8.69	13	7.14	0	0	0	0

Figure 3. Analysis of the peak topic *Visit of Pope*.

Virgin of Guadalupe became important because of the influence of the peak topic, while the inverse association reveals a possible influence of the peak topic over the topic *Raúl Salinas*, and indicates that the visit of Pope might have caused the oblivion of the sentencing of Raúl Salinas. Figure 3 illustrates these ephemeral associations.

6.3 Deviation detection

As yet another example, let us consider the first-page news from four Mexican newspapers, *El Universal*, *El Excelsior*[9], *El Financiero*[10], and *UnomásUno*, corresponding to the 4 of October of 1999. Table 2 lists the main topics for each one of these newspapers in this period.

The following newspapers are deviations: *El Universal*, since it did not mention the topic *UNAM* (acronyms of the National Autonomous University of Mexico)[11], and *UnomásUno* since it did not mention the topics *credits* and *Fox*[12].

Topics like *UNAM* and *Fox* seem to be the hot topics for that period, which is the reason for almost every newspaper to mention them. However, the topic *credits* is a possible interesting deviation since it is not a hot topic in the larger period,[13] but seems to be an important news topic that day. Then we can ask ourselves: What happened with *credits* (credits)? And also, why the newspaper *UnomásUno* did not mention it?

All newspapers have their own style: some prefer economic news (*El Financiero*, for instance), and others used to fill their front page with news coming from all sections. This causes many news reports to appear at the front page of only one newspaper; in spite of this, sometimes the topics appearing in only one of the newspapers are interesting deviations.

In our example there are, among some others, the following newspaper deviations.

El Excelsior mentioned topics like *Salinas* and *narcotics*; *UnomásUno* discussed *Labastida*[14]. With this, we may ask ourselves: Is there a new fact in the *Salinas* case? Why other newspapers did not mention it? Why *Labastida* appears in the first page of *UnomásUno*, and the other newspapers did not mention him?

7 Conclusions and future work

News reports are an important source of information about society. Their analysis allows to better understand its current interests and to measure the social importance of many events. In this paper, we presented a method for automated analysis of news topics. This method includes some of the basic functions of text mining: trend analysis, association discovery, and deviation detection.

In contrast with other text mining methods, our method is domain independent. This is an important characteristic for any method employed in the analysis of news reports.

We used only straight statistical measures in the discovering procedure, such as average, standard deviation, and correlation coefficient.

Some interesting features of our method are:

- We consider not only general change trends, but also stability trends. Other methods discover change trends (Lent *et al.*, 1997; Feldman & Dagan, 1995), but do not consider stability trends because in domains other than news, stability is not an important state.

[9] http://www.excelsior.com.mx

[10] http://www.financiero.com.mx

[11] This university was on strike during those days.

[12] Pre-candidate to the country presidency of PAN (National Action Party).

[13] We do not give here the analysis of the larger period.

[14] Pre-candidate to the country presidency of PRI (Institutional Revolutionary Party).

Newspapers	Topics
El Excelsior	*Alianza* (Alliance), *campaña* (campaign), *candidato* (candidate), *Colosio, conflicto* (conflict), *crédito* (credit), *elección* (election), *FOBAPROA, Fox, Madrazo, narcotráfico* (narcotics), *ONU, parista* (striker), *PRD, PRI, reforma* (reform), *Salinas, soberanía* (sovereignty), *UNAM, voto* (vote).
El Universal	*Acapulco, banco* (bank), *crédito* (credit), *elección* (election), *FOBAPROA, Fox, Madrazo, PRD, PRI, secuestro* (kidnapping), *violencia* (violence), *voto* (vote).
El Financiero	*Alianza* (Alliance), *banco* (bank), *Bush, campo* (countryside), *CGH, conflicto* (conflict), *crédito* (credit), *Fox, parista* (striker), *PAN, PRD, presupuesto* (budget estimate), *propuesta* (proposal), *sector agropecuario* (agriculture industry), *UNAM.*
UnomásUno	*Acapulco, asociaciones religiosas* (religious associations), *candidato* (candidate), *desastre* (disaster), *globalización* (globalization), *Guerrero, importaciones* (importations), *inflación* (inflation), *incremento* (increment), *Japón* (Japan), *Labastida, ONU, PRD, presupuesto* (budget estimate), *PRI, UNAM*

Table 2. Data for deviation detection

- We detect not only general trends, but also their factors (i.e. topics contributing in these trends).
- We discover not only associations indicating co-existence of topics (associations of the form $A \Rightarrow B$), but also associations expressing time relations between topics.
- We detect not only irregular topics, but also irregular collections (in this case, newspapers); for instance, newspapers that did not mention something considered important by many others.

Finally, it is important to point out that such discovery of facts from news reports (trends, associations and deviations) helps to interpret the social importance of news topics. In addition, it allows finding some parameters for an improved characterization of news reports and of society interests.

As future work we plan to focus on the following two tasks:

- Improve the extraction of the news topics. Here, we plan to use recent information extraction and text categorization techniques (Gelbukh *et al*., 1999).
- Increase the kind of the analysis operations. We plan to design a method for the conceptual clustering of the news reports, and to use this clustering in the construction of a news summary. Additionally, we plan to study the classification favorable and unfavorable news (García-Menier, 1998).

References

Ah-Hwee Tan, Text Mining: The state of the art and challenges, *Proc. of the Workshop Knowledge Discovery from advanced Databases PAKDDD-99*, April 1999.

Ahonen-Myka, Helena, **Oskari Heinonen**, **Mika Klemettinen**, and **A. Inkeri Verkamo**, Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery, *Proc. of 16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, 1999.

Arning Andreas, **Rakesh Agrawal**, and **Prabhakar Raghavan**, A Linear Method for Deviation Detection in Large Databases, *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining*, 1996.

Cramér, Harald. *Mathematical Methods of Statistics*, Landmarks in Mathematics and Physics, Princeton, 1999.

Fayyad, Usama M., **Gregory Piatetsky-Shapiro**, **Padhraic Smyth**, and **Ramasamy Uthurusamy**, *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 1996.

Feldman R (editor), Proc. of The 16th International Joint Conference on Artificial Intelligence, *Workshop on Text Mining: Foundations, Techniques and Applications*, Stockholm, Sweden, 1999.

Feldman R., and **I. Dagan**, Knowledge Discovery in Textual Databases (KDT), *Proc. of the First International Conference on Knowledge discovery (KDD_95)*, pp. 112-117, Montreal, 1995.

Feldman, R., and **H. Hirsh**, Mining Associations in Text in the Presence of Background Knowledge, *Proc. of the 2nd International Conference on Knowledge Discovery (KDD-96)*, pp. 343-346, Portland, 1996.

Feldman, R., **M. Fresko**, **Y. Kinar**, **Y. Lindell**, **O. Liphstat**, **M. Rajman**, **Y. Schler**, and **O. Zamir**, Text Mining at the Term Level, *Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, Nantes, France, September 23-26, 1998.

Freund and **Walpole**, *Estadística Matemática con Aplicaciones*, Cuarta Edición, Prentice Hall, 1990.

García-Menier, E., Un Sistema para la Clasificación de Notas Periodisticas, *Proc. of the Simposium Internacional de Computación CIC-98,* México, D. F., 1998.

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

Gelbukh, A., G. Sidorov, A. Guzmán-Arenas. Use of a weighted topic hierarchy for document classification**.** *Proc. 2nd International Workshop TSD-99***,** Plzen, Czech Republic, September 13-17, 1999.

Glymour, C., **D. Madigan**, **D. Pregibon**, **P. Smyth**, Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery* 1, 11-28, 1997.

Han and **Kamber,** *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001

Hearst, Marti A., Untangling Text Data Mining, *Proc. of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Marylnd, June 20-26, 1999.

Knorr, E., **R. Ng**, and **V. Tucakov**. Distance-Based Outliers: Algorithms and Applications, *The VLDB Journal*, 8(3): 237-253, 2000.

Lent, Brian, **R. Agrawal,** and **R. 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.

Mladenic Dunja, Proc. of the Sixth International Conference on Knowledge Discovery and Data Mining, *Workshop on Text Mining*, Boston, MA, 2000.

Montes-y-Gómez, M., **A. Gelbukh**, **A. López-López,** Discovering Ephemeral Associations among news topics, *17th International Joint Conference on Artificial Intelligence IJCAI-01, Workshop on Adaptive Text Mining*, 2001.

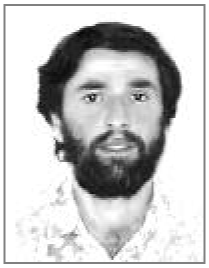
Montes-y-Gómez, M., **A. Gelbukh**, **A. López-López**, and **R. Baeza-Yates**, Flexible Comparison of Conceptual Graphs, *To appear in the Proc. of the 12th International Conference and Workshop on Data-base and Expert Systems Applications, DEXA-2001*, Springer, 2001.

Nahm and **Mooney**, A Mutually Beneficial Integration of Data Mining and Information Extraction, *Proc. of the Seventeenth Conference of Artificial Intelligence, AAAI-2000*, Austin, TX, 2000.

Rajman, Martin, and **Romaric Besançon**, Text Mining – Knowledge Extraction from Unstructured Textual Data, *6th Conference of International Federation of Classification Societies (IFCS-98)*, 473-480, Rome, July 21-24, 1998.

Shian-Hua Lin, **Chi-Sheng Shin**, **Meng Chang Chen**, **Jan-Ming Ho**, **Ming-Tat Ko**, and **Yueh-Ming Huang**, Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A semantic Approach, *Proceedings of SIGIR'98*, Melbourne, Australia, 1998.

Manuel Montes y Gómez *received the B.Sc. degree in Electronics (1996) from the Technological Institute of Morelia, Mexico, and the M.Sc. degree in Electronics (1998) from the National Institute for Astrophysics, Optics and Electronics (INAOE), Puebla, Mexico. Currently he is a Ph.D. student of the Natural Language Laboratory of the Center for Computing Research (CIC), IPN, Mexico. His main research interests are: text mining, information extraction and retrieval, and statistical natural language processing.*



Alexander Gelbukh *received his M.Sc. degree in mathematics (1990) from the Moscow State "Lomonossov" University, Russia, and his Ph.D. degree in Computer Science (1995) from VINITI, Russia. Currently he is a Research Professor and the head of the Natural Language Laboratory of the Center for Computing Research (CIC), IPN, Mexico. He is a member of Mexican Academy of Sciences and of National System of Researchers (SNI) of Mexico. His main research interests include computational linguistics, natural language processing, and information retrieval. See http://www.cic.ipn.mx/~gelbukh.*

Aurelio López López *is a professor at the Computational Sciences Department of the INAOE, Tonantzintla, Puebla, México. He got his PhD in Computer and Information Science from Syracuse University, Syracuse NY, U.S.A. His areas of interest include knowledge representation, information extraction and retrieval, natural language processing, and text mining.*