Detecting the Dependencies of a Peak News Topic[*]

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

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

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

Abstract

News topics in a society are volatile and face a constant change, but in their lifetime, while they drop, interact with other topics and influence them. In this paper, we describe the dependencies between peak news topics and other current topics, and propose a procedure to detect them. We also explain how these "observable" dependencies (discovered from the newspapers) are not always "real world" dependencies, and how we can infer real ones from them. Finally, we argue that the discovery of these dependencies can be translated into knowledge about interests of society and social behavior.

Keywords: *text mining, data mining, natural language processing, news, and statistics.*

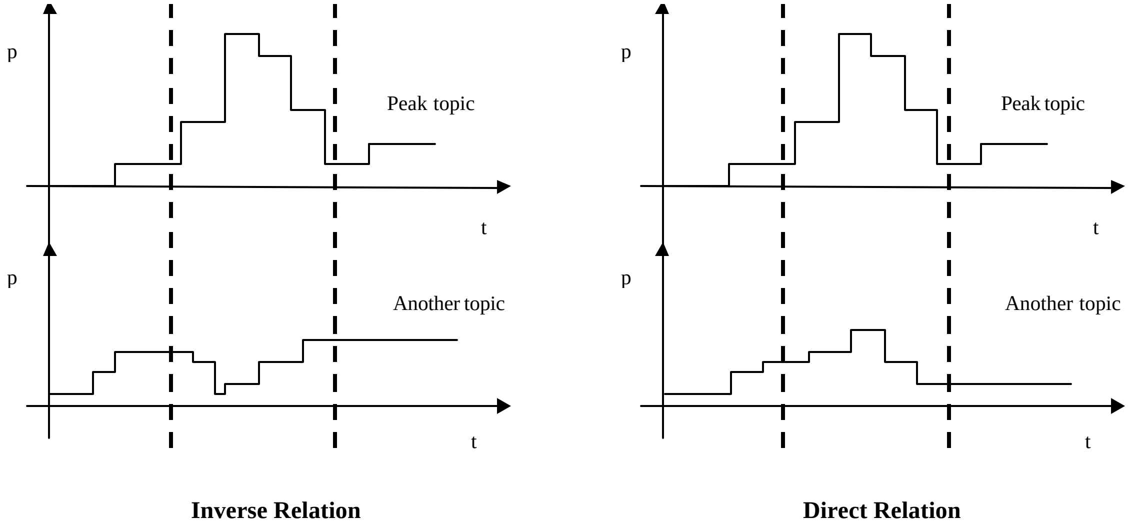
1 Introduction

One of the newest research areas of text processing is *text mining*. Text mining is, in essence, the discovering of new knowledge from texts. Some of the important tasks tackled by text mining are: trend analysis, document clustering, deviation detection and discovery of rules of association [Hearst, 1999; Montes-y-Gómez et al., 1999; García-Menier, 1998; Lent et al., 1997; Feldman & Dagan, 1995].

Within the discovery of rules of association, the objective is to uncover hidden associations occurring between concepts used in texts. For instance, for a medical corpus, an association between 'cancer' and 'therapy' can emerge.

In this direction, we present a method for analyzing news with the aim of discovering a special kind of association between the topics reported over a time span. We are interested in topics with one-time short-term peaks of frequency, i.e., such that their importance sharply rises within a short period and very soon disappears; for instance, the visit of Pope John Paul II to Mexico City. We describe a technique that discovers the influence of such topics on other current topics reported on the news.

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



Inverse Relation

Direct Relation

Figure 1. Types of dependencies between opinion topics.

Based on the fact that topics with such sharp peaks distract the attention of people from other topics, causing momentary oblivion of others themes, we focus on the identification of linear (or quasi-linear) dependencies between those "sharp" topics and other news topics. By linear dependency between two graphics f_1 and f_2 we mean that there exists some constants a and b such that $f_1 = a f_2 + b$.

Figure 1 illustrates these ideas and shows the inverse and direct dependencies that can occur between two topics.

For the analysis of these topics and the discovery of their relations, we represent topics by their probability distributions and use the correlation measure r for identifying dependencies between them.

2 Dependencies between topics

It is important to distinguish two different statistical characteristics of the topics appearing in the newspapers. One characteristic is the "real world" frequency: the frequency with which the corresponding news come from information agencies, for instance. Another characteristic is the "observable" frequency, expressed as the pieces of news actually appearing in the newspapers.

To illustrate the difference, let us consider two sources of information: say, a journalist working in Columbia and another one working in Kosovo. Let the first one sends 30 messages each month and the second one sent 30 messages in January and 70 messages in April 1999. These are the "real world" frequencies: 30 and 30 in January, and 30 and 70 in April. However, the newspaper has a fixed size and can only publish, say, 10 messages per day. Then it will publish 5 and 5 messages from these correspondents in January, but 3 and 7 in April. These are the "observable" frequencies, since this is the only information we have from the newspaper texts.

Thus, the observable frequencies can be considered normalized (their sum is a constant) while the "real" ones are not normalized. In our model, these two frequencies are proportional, but only one of them, namely, the observable frequency, is normalized, the pro-

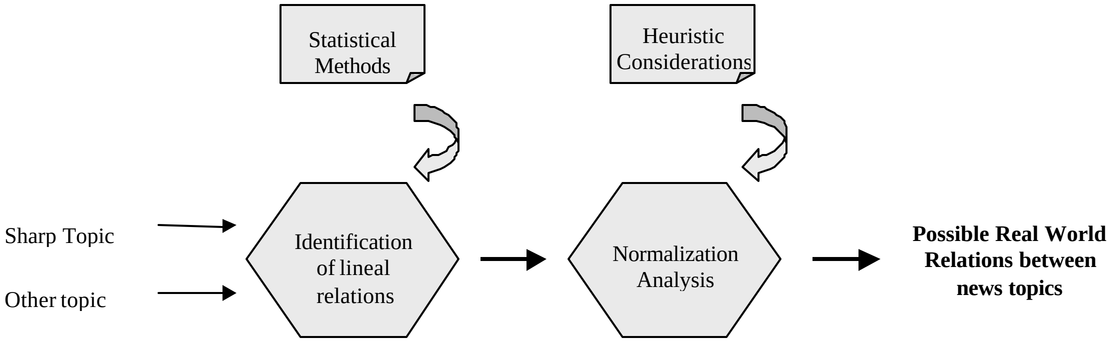


Figure 2. Method of discovering dependencies between topics.

portion coefficient being the normalization constant.

We define a *real* topic dependency as a dependency that holds between the topics in the real word and not only in our graph representing the (normalized) observable frequencies. Thus, a graphical dependency is a combination of two sources: a possible real-world dependency and the normalization.

3 Discovering the "real world" dependencies

Given a peak topic and the observable data surrounded, our task is to determine whether, a real dependency holds between the peak topic and other topic and the kind of dependency. For determining the absence or presence of dependency, we propose the following procedure:

1. Detect dependencies in the observed data;
2. Determine whether the dependency is due only to normalization or there is a *possible* real-world dependency component.

Let us consider these steps in more detail.

In the first step, by means of statistical analysis of the probability distributions of topics we detect the groups of topics having a graphical (observable) dependency and thus *candidates* to have a real topic dependency. The statistical method we use to detect this observable dependency is the correlation measure r [Freund & Walpole, 1990], expressed as:

$$r = \frac{S_{01}}{\sqrt{S_0 S_1}}$$

where:

$$S_0 = \sum \left(p_i^0\right)^2 - \frac{1}{n}\left(\sum p_i^0\right)^2$$

$$S_1 = \sum \left(p_i^1\right)^2 - \frac{1}{n}\left(\sum p_i^1\right)^2$$

$$S_{01} = \left(\sum p_i^0 \; p_i^1\right) - \frac{1}{n}\left(\sum p_i^0\right)\left(\sum p_i^1\right)$$

Here p^0 are the probabilities of the peak topic and p^1 are the probabilities of the topic in question.

The correlation coefficient r measure how well two variables are related to each other. 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 that exist an exact direct relation between the variables

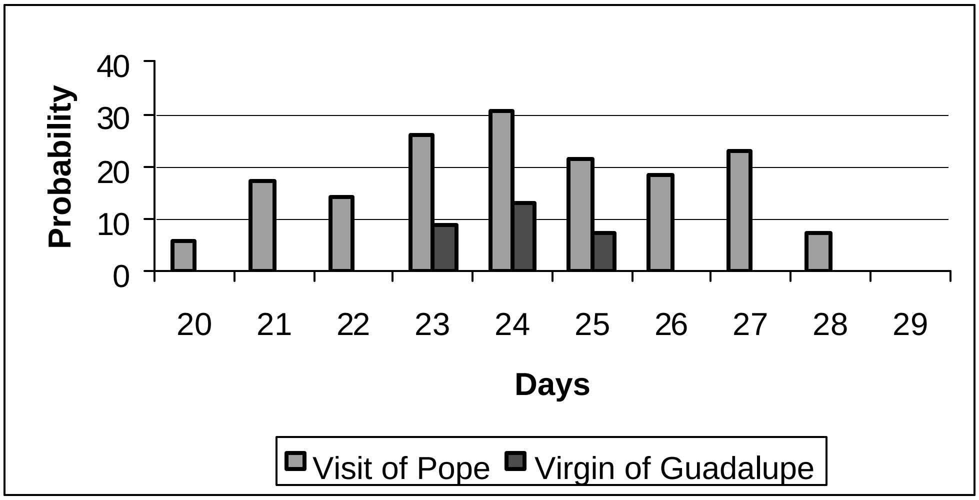


Figure 3. Dependency between the topics
Visit of Pope and *Virgin of Guadalupe*

Source data

Time	p_{peak}	$p_{i\ other}$
20 January	5.53	0
21 January	17.39	0
22 January	14.28	0
23 January	26.08	8.69
24 January	30.43	13.04
25 January	21.42	7.14
26 January	18.18	0
27 January	23.07	0
28 January	7.14	0
29 January	0	0

Correlation coefficient

Period	*r*
23 to 25 of January	0.959
20 to 29 of January	0.718

Table 1. Data and correlation coefficient between the topics
Visit of Pope and *Virgin of Guadalupe*

and 0 that that do not exist any relation between the variables at all.

In the second step, we first analyze the type of dependency found. If we found a direct dependency between the topics, then this dependency has a high probability for being a "real world" one. If we found an inverse dependency, we need to check whether the hypothesis of pure normalization dependency is satisfied. If it is, then the observed relation between the change of the peak theme and the change of the given topic does not hold in the real world frequencies.

Let us formulate more precisely the hypothesis of real world independence:

- There is a purely normalization dependency between the two topics, i.e., no real-world dependency;

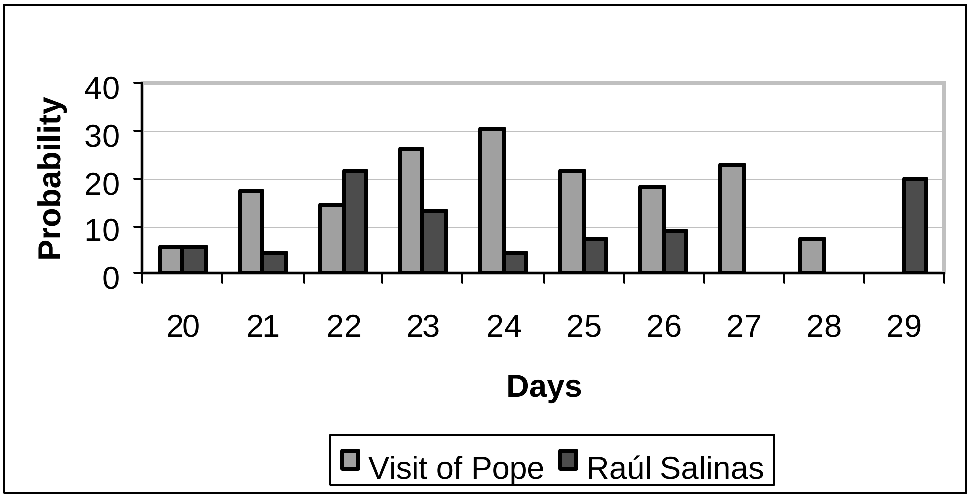


Figure 4. Observable dependency between the topics *Visit of Pope* and *Raúl Salinas*

Source data

Time	P_{peak}	$p_{i\ other}$
20 January	5.53	5.55
21 January	17.39	4.34
22 January	14.28	21.42
23 January	26.08	13.04
24 January	30.43	4.34
25 January	21.42	7.14
26 January	18.18	9.09
27 January	23.07	0
28 January	7.14	0
29 January	0	20

Correlation coefficient

Period	*r*
22 to 25 of January	–0.818
22 to 26 of January	–0.703
22 to 27 of January	–0.601
21 to 27 of January	–0.462

Table 2. Data and correlation coefficient between the topics *Visit of Pope* and *Raúl Salinas*

- No other topics correlate with the peak one;
- There are no significant real-world changes in this period in all topics except for the peak one.

Then the proportions between the given topic and all the topics co-occurring with the peak (those appearing during all the period of existence of the peak) are the same before and during the peak: $p^1_j \,/\, p^i_j = p^1_{j+1} \,/\, p^i_{j+1}$, with p^i_j representing the probability of topic *i* in time *j*, p^1 the probability of the given topic, *i* any topic that co-occurs with the peak one, and *j* the time spans corresponding to the peak.

The probability for this equation to hold just by accident is very low. Thus, this equation defines a high-probable criterion of the absence of any real world dependency between the peak topic and any of the other topics.

When the conditions above do not hold, and as a consequence the equation do not hold either, the correlation coefficient r is used as a second criterion of the probability of any real dependency between the topics in question.

The interpretation of the correlation coefficient in this context is that when $r = -1$ exist an exact linear dependency between the topics and a high probability that the dependency is caused by pure normalization. Thus, when $r \to -1$ the probability of a dependency caused by pure normalization increases, and when $r > -1$ and $r << 0$ the probability of a "real world" dependency growth. Figure 2 shows the main components of our method.

4 Experimental results

To test these ideas, we used the Mexican newspaper *El Universal*. We collected the national news for the ten days surrounding the visit of Pope John Paul II to Mexico City, i.e. from January 20 to 29 of 1999, and looked for some dependency between the peak topic and the other topics.

Before analyzing the data, we first identified all news topics using a method similar to that proposed by Gay & Croft [1990], where the topics are related to noun strings. Then we constructed the probability distribution for each one of the topics, including the peak topic (*the visit of Pope*).

To discover some dependency with the peak topic, we calculated the correlation coefficient r between each one of the topics and the peak topic, and then we applied the criteria described above to select the *possible* real-world dependencies.

One of the dependencies detected with our method was a direct dependency between the peak topic, *the visit of Pope*, and the topic *Virgin of Guadeloupe*. Figure 3 and Table 1 illustrate this relation.

Another interesting discovery was the inverse dependency between the peak topic, *the visit of Pope*, and the topic *Raúl Salinas* (Brother of the Mexican ex-president, Carlos Salinas de Gortari, sentenced in those days). Figure 4 shows this dependency and table 2 displays the data and the correlation coefficient.

These results indicate that in the period from the 21 to the 27 of January there was an observable dependency between the peak topic, *Visit of pope*, and the topic *Raúl Salinas*.

The hypothesis of pure normalization did not hold and the values of the correlation coefficient, $r = -0.462$, $r = -0.601$, $r = -0.703$ and $r = -0.818$ satisfied the conditions $r > -1$ and $r << 0$. Thus, there is a high possibility that this dependency was a real world one. If this is true, then we can say that the topic *Raúl Salinas* went out of the attention because of the influence of the peak topic.

5 Conclusions and future work

We have described a phenomenon we think is very frequent in real life situations; the influence of peak topics over other news topics. By peak topics, we mean those whose importance sharply rises within a short period of time and very soon sharply drops: for instance, the visit of Pope John Paul II to Mexico City.

For the analysis of these relations, we distinguish between two types of dependencies: "real world" and observable dependencies (subjects of our analysis). We have proposed a model in which "real world" dependencies can be inferred from observable dependencies.

As future work, we plan to test these ideas and criteria under different situations and to use them to detect special circumstances (favorable scenarios and difficult conditions) that make the discovering process more robust and exact.

Finally, its important to point out that the discovery of this kind of dependencies helps to interpret the social importance and influence of relevant but transitory topics and to define some parameters for an improved characterization of them and of society.

References

[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.

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

[García-Menier, 1998] Everardo García Menier, Un sistema para la Clasificación de notas periodisticas, Proc. Of the Simposium Internacional de Computación, 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.

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

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

[Lent et al., 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.

[Montes-y-Gómez et al, 1999] M. Montes-y-Gómez, A. López-López, A. Gelbukh. Text Mining as a Social Thermometer. Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, August 1999.