A Statistical Approach to the Discovery of Ephemeral Associations among News Topics*

M. Montes-y-Gómez [1], A. Gelbukh [1], A. López-López [2]

[1] Center for Computing Research (CIC), National Polytechnic Institute (IPN), 07738, Mexico.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

[2] Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), Puebla, Mexico.
allopez@inaoep.mx

Abstract. News reports are an important source of information about society. Their analysis allows understanding its current interests and measuring the social importance and influence of different events. In this paper, we use the analysis of news as a means to explore the society interests. We focus on the study of a very common phenomenon of news: the influence of the peak news topics on other current news topics. We propose a simple, statistical text mining method to analyze such influences. We differentiate between the *observable* associations—those discovered from the newspapers—and the *real-world* associations, and propose a technique in which the real ones can be inferred from the observable ones. We illustrate the method with some results obtained from preliminary experiments and argue that the discovery of the ephemeral associations can be translated into knowledge about interests of society and social behavior.

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 fixed structure, such as databases with fields having no complex structure of their own. The methods for the analysis of large databases and the discovery of new knowledge from them 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 the discovery of new facts and world knowledge from large collections of texts that—unlike those considered in 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 collection of texts.

In this paper, we focus on the analysis of a collection of news reports appearing in newspapers, newswires, or other mass media. The 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

* Work done under partial support of CONACyT, CGEPI-IPN, and SNI, Mexico.

society interests and behavior, they are very diverse and constantly changing, and also they interact with, and influence, each other.

Some previous methods consider: the trend analysis of news (Montes-y-Gómez *et al*., 1999), the detection of new events on a news stream (Allan *et al*., 1998), and the classification of bad and good news (García-Menier, 1998). Here, we focus on the analysis of a very common phenomenon of news: the influence of the peak news topics over other current news topics.

We define a peak news topic as a topic with one-time short-term peak of frequency of occurrence, i.e., such that its importance sharply rises within a short period and very soon disappears. For instance, the visit of Pope John Paul II to Mexico City became a frequent topic in Mexican newspapers when the Pope arrived to Mexico and disappeared from the newspapers in a few days, as soon as he left the country; thus this is a peak topic.

Usually, these topics influence over the other news topics in two main ways: a news topic induces other topics to emerge or become important along with it, or it causes momentary oblivion of other topics.

The method we proposed analyzes the news over a fixed time span and discovers just this kind of influences, which we call *ephemeral associations*.

Basically, this method uses simple statistical representations for the news reports (frequencies and probability distributions) and simple statistical measures (the correlation coefficient) for the analysis and discovery of the ephemeral associations between news topics (Glymour *et al*., 1997).

Additionally, we differentiate between the *observable* ephemeral associations, those immediately measured by the analysis of the newspapers, and the *real-world* associations. In our model, the real-world associations in some cases can be inferred from the observable ones, i.e., for some observable associations its possibility to be a real-world one is estimated.

The rest of the paper is organized as follows. Section 2 defines ephemeral associations and describes the method for their detection. Section 3 introduces the distinction between the observable and the real-world associations and describes the general algorithm for the discovery of the real-world associations. Section 5 presents some experimental results. Finally, section 6 discusses some conclusions.

2 Discovery of Ephemeral Associations

A common phenomenon in news is the influence of a peak 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.[1] An ephemeral association can be viewed as a direct or inverse relation between the prob-

[1] This kind of associations is different from the associations of the form $X \Rightarrow Y$, because they not only indicate the co-existence 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.

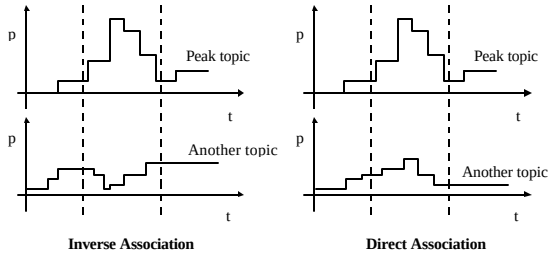

Figure 1. Ephemeral associations between news topics

ability distributions of the given topics over a fixed time span. Figure 1 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 other topic, while an inverse ephemeral association suggests that the peak topic probably produces the momentary oblivion of the other news topic.

Thus, given a peak topic and the surrounding data, we detect the ephemeral associations in two steps:

1. Construction of the probability distribution for each news topic over the time span around the peak topic.
2. Detection of the ephemeral associations in the observed data set (if any).

These steps are described in the next two subsections.

2.1 Construction of the Probability Distributions

Given a collection of news reports corresponding to the time span of interest, i.e., the period around the existence of the peak topic, we construct a structured representation of each news report, which in our case is 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 proper to Spanish and based on proximity of words that allow identifying and extracting phrases. These rules are guided by the occurrence of articles and some times by the occurrence of the prepositions *de* or *del* (of in English) along with nouns or proper nouns. For instance, given the following paragraph, the highlighted words are selected as keywords.

"La demanda de acción de inconstitucionalidad tiene como argumentos una serie de violaciones que el Congreso de Yucatán incurrió porque, de acuerdo con el PRD, hizo modificaciones a la ley electoral 90 días antes de que se lleven a cabo los comicios en ese Estado de la República".

Once this procedure is done, a frequency f_k^i can be assigned to each news topic. The frequency f_k^i is calculated as the number of news reports for the day *i* that men-

tion the topic k. It is more convenient, however, to describe each news topic k by a probability distribution $D_k = \{p_k^i\}$ by the days i, where for a given day i, p_k^i expresses the probability for a news topic randomly chosen from the reports of that day to be the topic k:[2]

$$p_k^i = f_k^i \Big/ \sum_{j \in Topics} f_j^i \qquad (1)$$

We will call the values p_k^i *relative probabilities*. A probability $p_k^i = 0$ indicates that the topic k was not mentioned on the day i. The relative probabilities p_k^i are advantageous for the discovery of the ephemeral associations mainly because they maintain a normalization effect over the news topics: for any day i,

$$\sum_{k \in Topics} p_k^i = 1$$

This condition holds for the whole period of interest and means that the increase of the relative probability of one news topic always is compensated by the decrease of probability of some other topics, and vice versa.

2.2 Detection of Ephemeral Associations

The ephemeral associations express inverse or direct relations between the peak topic and some other current news topics (see the figure 1). Let the peak topic be, say, the topic $k = 0$ and the other one we are interested in be, say, the topic $k = 1$. The statistical method we use to detect the observable associations is based on the correlation measure r between the topics $k = 0$ and $k = 1$(Freund and Walpole, 1990) defined as:

$$r = \frac{S_{01}}{\sqrt{S_{00} S_{11}}}, \quad \text{where} \quad S_{kl} = \sum_{i=1}^{m} \left(p_k^i p_l^i\right) - \frac{1}{m}\left(\sum_{i=1}^{m} p_k^i\right)\left(\sum_{i=1}^{m} p_l^i\right), \quad k,l = 0,1. \qquad (2)$$

Here, p_k^i are defined in the previous section and m is the number of days of the period of interest.

The correlation coefficient r measures how well the two news topics are related to each other.[3] Its values are between –1 and 1, where –1 indicates that there exists an exact inverse relation between the two news topics; 1 indicates the existence of an exact direct relation between the news topics, and 0 the absence of any relation at all.

Therefore, if the correlation coefficient between the peak topic and some other news topic is greater than a user-specified threshold u (i.e., $r < -u$) then there exists a *direct* ephemeral association between them. On the other hand, if the correlation

---

[2] This roughly corresponds to the percentage of the space the newspapers devoted to the topic k on the day i.

[3] The usual interpretation of the correlation coefficient is the following: $100\ r^2$ 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.

coefficient is less than the threshold $-u$ (i.e., $r < -u$) then there exists an *inverse* ephemeral association between the two topics.

There are two reasons for introducing the user-specified threshold u. First, it softens the criterion so that we can approximate the way a human visually detects the association. Second, to cope with the relatively small data sets in our application: since few data are available (a peak topic persists over few days), random variations of topic frequencies unrelated to the effect in question can greatly affect the value of the correlation coefficient. A typical value recommended for u is around 0.5.

3 Discovery of Real World Associations

Newspapers usually have a fixed size, and the editor has to decide what news to include in the day's number and what not to include. Thus, the frequencies of the news mentioned in a newspaper do not directly correspond to the number of events that happen in the real world on the same day.

In the next subsection, we explain this important difference between the real world news frequencies and the ones normalized by the fixed size of the newspaper. Then, we show how to estimate whether the observable associations are mainly due to the normalization effect or there is a possible real-world association component.

3.1 The Notion of Real World Associations

Since our ultimate goal is to discover the associations that hold in the real world, it is important to distinguish between 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 comes from the information agencies, for instance. Another characteristic is the observable frequency, expressed as the pieces of news actually appearing in the newspapers.

To illustrate this difference, let us consider two sources of information: say, a journalist working in Colombia and another one working in Salvador. Let the first one sends 30 messages each week and the second one sent 30 messages in the first week and 70 messages in the second week. These are the real-world frequencies: 30 and 30 in the first week, and 30 and 70 in the second one (i.e., there was something interesting happening in Salvador in the second week). However, the newspaper has a fixed size and can only publish, say, 10 messages per week. Then it will publish 5 and 5 messages from these correspondents in the first week, but 3 and 7 in the second week. These are the observable frequencies, since this is the only information we have from the newspaper texts.

Our further considerations are based on the following two assumptions.

Assumption 1: The newspapers tend to have a constant "size."[4]

Thus, the observable frequencies can be considered normalized, i.e., their sum is a constant, while the real world ones are not normalized. We assume that these two

[4] The "size" of a newspaper not only depends on its physical size (for instance, the number of pages) but also on the number of the journalists, the time required for editing, printing, etc.

kinds of frequencies are proportional, being the proportion coefficient the normalization constant. Thus, we define a real-world ephemeral association as an association that holds between the topics in the real world and not only in the observable (normalized) data, and we consider that an observable ephemeral association is a combination of two sources: a (possible) real-world ephemeral association and the normalization.

The normalization effect is always an inverse correlation effect. This means that the increase of probability of the peak topic is always compensated by the decrease of probability of some other topics, and vice versa. Thus, we can conclude that any *direct* observable ephemeral association is, very probably, a real-world association.

Assumption 2: The peak topic proportionally takes away some space from each current news topic.

First, this assumption implies that the relative proportions among the rest of the news topics do not change if we take away the peak topic and its related topics. Second, no topic completely disappears only as a consequence of the normalization effect.[5]

3.2 Detection of Real World Associations

As we have noted, all direct associations should be considered real-world ones, so we only need to analyze the inverse ones. The idea is to restore the original distribution of the topics by eliminating the normalization effect, and check if this distribution still correlates with that of the peak topic.

The Assumption 2 allows us to estimate the probability distribution $D'_k = \{p'^i_k\}$ of the topic k as it would be without the normalization effect, where the probability p'^i_k expresses the relative probability of occurrence of the topic k on the day i after we take away the peak topic and it related topics. This probability is calculated as follows:

$$p'^i_k = f^i_k \Big/ \sum_{j \notin Peak} f^i_j \qquad (3)$$

Here the set *Peak* consists of the peak topic and its related topics (those with a direct association), while the frequency f^i_k indicates the number of news reports in the day i that mention the topic k.

Therefore, an inverse observable association between the peak topic and the news topic k is likely a real-world association if it remains after the normalization effect is eliminated from the topic k. In other words, if the correlation coefficient between the peak distribution and the corrected distribution D'_k is less than the user-specified threshold $-u$ (i.e., $r < -u$) then the inverse observable ephemeral association is likely a real-world one.

[5] Usually, newspaper editors design the newspaper format and contents in such a way that they expose all news of the day, even if briefly.

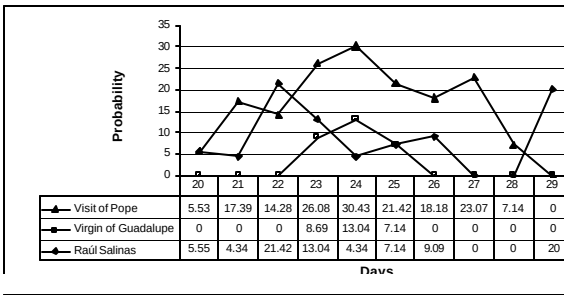

Figure 2. Analysis of the peak topic "*Visit of Pope*"

Concluding, the our basic algorithm for the discovery of the real-world ephemeral associations among the news topics of a given period consists of the following steps:

1. Calculate the (observable) probabilities by the formula (1);
2. Calculate the (observable) correlations between the peak topic and the other ones by the formula (2);
3. Select the topics that strongly correlate with the peak one, using a threshold $u \approx 0.5$;
4. Determine which associations are real-world ones:
   a. All direct associations are real-world ones;
   b. For the inverse associations,
      i. Build the corrected distributions by the formula (3), using the knowledge obtained at the step 3;
      ii. Calculate the (real-world) correlations between the peak topic and the other ones using the formula (2) and the corrected distributions;
      iii. The topics for which this correlation is strong represent the real-world associations.

4 Experimental Results

To test these ideas, we used the Mexican newspaper *El Universal*.[6] 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, 1999, and looked for some ephemeral associations between this peak topic and the other topics.

One of the associations detected with our method (using the threshold $u = 0.6$) was a direct ephemeral association between the peak topic and the topic *Virgin of Guadalupe.*[7] The figure 2 illustrates this association. The correlation coefficient was

[6] http://www.el-universal.com.mx
[7] A Mexican saint whose temple the Pope visited.

$r = 0.959$ for the period between the 23 and 25 of January (stay of the Pope in Mexico), and $r = 0.719$ for the surrounding period between the 20 and 29 of January.

Since this association was a direct one, it had a high possibility for being a real-world one. This means that the topic *Virgin of Guadalupe* probably emerged because of the influence of the peak topic. Moreover, since this topic was the only one that had a direct association with the peak topic, we deduced that the visit of the Pope was strongly related with *Virgin of Guadalupe* (in fact, he has focused his discourse on this important Mexican saint).

Another interesting discovery was the inverse association between the peak topic and the topic *Raúl Salinas* (brother of the Mexican ex-president, Carlos Salinas de Gortari, sentenced in the 22 of January). The figure 2 also shows this association.

The correlation coefficient $r = -0.703$ between the 22 and 26 of January (period covering the *Visit of Pope* and the sentencing of *Raúl Salinas*) indicates the existence of an inverse observable ephemeral association.

In order to determine the possibility of this association for being a real-world one, we analyzed the normalization effect. First, we built the probability distribution of the topic *Raúl Salinas* without considering the peak topic and its related topics (the topic *Virgin of Guadalupe* in this case). The new probability distribution was:

$$D'_{RaulSalinas} = \{5.88,\ 5.26,\ 25,\ 17.64,\ 6.25,\ 9.09,\ 11.11,\ 0,\ 0,\ 20\}$$

Second, we recomputed the correlation coefficient between the peak topic and the topic *Raúl Salinas*. The new correlation coefficient $r = -0.633$ (between the 22 and 26 of January) indicated that it was very possible for this association to be real-world one. If this was true, then the topic *Raúl Salinas* went out of the attention because of the influence of the visit of the Pope to Mexico City.

As another example, we examined the peak topic *Death of Kennedy Jr*. This topic took place between the 18 and 24 of July of 1999. For the analysis of this peak topic, we used the news appearing in the national desk section of the newspaper *The New York Times*.[8] Among our discoveries, there were two inverse ephemeral associations. One of them between the peak topic and the topic *Election 2000*, with $r = -0.68$, and the other one between the peak topic and the topic *Democrats*, with $r = -0.83$. The figure 3 shows these associations.

Since these associations were both inverse ones, we analyzed their normalization effect. First, we built their probability distributions without considering the peak topic:

$$D'_{Election2000} = \{0,\ 9.52,\ 0,\ 5.26,\ 0,\ 0,\ 0,\ 2.94,\ 11.11\}$$
$$D'_{Democrats} = \{11.53,\ 4.76,\ 0,\ 0,\ 0,\ 0,\ 0,\ 2.94,\ 7.4\}$$

Then, we recomputed their correlation coefficients. The probability distribution of the topic *Democrats* did not change (because of the zero probabilities of the topic

[8] The topics were extracted manually as opposed to the Spanish examples that were analyzed automatically.

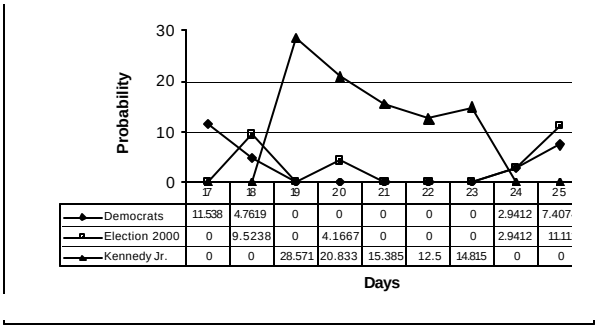

Figure 3. Analysis of the peak topic "*Death of Kennedy Jr.*"

Democrats during the peak existence). Thus, the correlation coefficient was again $r = -0.83$ and we concluded that this association had a high possibility for being a real-world one.

On the other hand, the new correlation coefficient between the topic *Elections 2000* and the peak topic, $r = -0.534$, was not less than the threshold $-u$ (we used $u = 0.6$), therefore, there was not enough evidence for this association to be a real-world one.

5 Conclusions

We have analyzed a very frequent phenomenon in real life situations—the influence of a peak news topic on the other news topics. We have described a method for the discovery of this type of influences, which we explain as a kind of association between the two news topics and call *ephemeral associations*. The ephemeral associations extend the concept of typical associations because they not only reveal co-existence relations between the topics but also their temporal relations.

We distinguish between two types of ephemeral associations: the observable ephemeral associations, those discovered directly from the newspapers, and the real-world associations. We have proposed a technique with which the observable associations are detected by simple statistical methods (such as the correlation coefficient) and the real-world associations are heuristically estimated from the observable ones.

For the sources that do not have any fixed size, such as newswires, the observed frequencies of the news reports correspond to the real world ones. For such sources, the method discussed in this paper do not make sense. An easier way to discover the same associations in this case is not to normalize the frequencies in the formula (1), using $p_k^i = f_k^i$ instead and then applying the formula (2).

However, if it is not clear or not known whether the news source presents the normalization problem, then the method presented here can be applied indiscriminately.

This is because in the absence of normalization effect, our method will give equally correct results, though with more calculations.

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 precise. Basically, we plan to experiment with multiple sources, and to analyze the way their information can be combined in order to increase the precision of the results.

Finally, it is important to point out that the discovery of this kind of associations, the ephemeral associations among news topics, helps to interpret the relationships between society interests and discover hidden information about the relationships between the events in social life.

References

1. Ahonen-Myka, Heinonen, Klemettinen, and Verkamo (1999), Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Stockholm, 1999.
2. Allan, Papka and Lavrenko (1998), On-line new Event Detection and Tracking, Proc. of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, August 1998
3. Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996), Advances in Knowledge Discovery and Data Mining, Cambridge, MA: MIT Press, 1996.
4. Feldman, editor (1999), Proc. of The 16th International Joint Conference on Artificial Intelligence, Workshop on Text Mining: Foundations, Techniques and Applications, Stockholm, Sweden, 1999.
5. Feldman and Hirsh (1996), Mining Associations in Text in the Presence of Background Knowledge, Proc. of the 2nd International Conference on Knowledge Discovery (KDD-96), Portland, 1996.
6. Freund and Walpole (1990), Estadística Matemática con Aplicaciones, Cuarta Edición, Prentice Hall, 1990. (In Spanish)
7. García-Menier (1998), Un sistema para la clasificación de notas periodísticas, Simposium Internacional de Computación CIC-98, México, D.F., 1998.
8. Gay and Croft (1990), Interpreting Nominal Compounds for Information Retrieval, Information Processing and Management 26(1): 21-38, 1990.
9. Glymour, Madigan, Pregibon, and Smyth (1997), Statistical Themes and Lessons for Data Mining. Data Mining and Knowledge Discovery 1, 11-28, 1997.
10. Han and Kamber (2000), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
11. Hearst (1999), Untangling Text Data Mining, Proc. of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Marylnd, 1999.
12. Mladenic (2000), Proc. of the Sixth International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining, Boston, MA, 2000.
13. Montes-y-Gómez, López-López and Gelbukh (1999), Text Mining as a Social Thermometer, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Stockholm, 1999.
14. Rajman and Besançon (1998), Text Mining - Knowledge Extraction from Unstructured Textual Data, 6th Conference of International Federation of Classification Societies (IFCS-98), Rome, 1998.