Finding Correlative Associations among News Topics

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

[1] Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan Dios Bátiz s/n esq. Mendicabal, col. Zacatenco, CP. 07738, DF, Mexico.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

[2] INAOE, Luis Enrique Erro No. 1, Tonantzintla, Puebla, 72840 México.
allopez@inaoep.mx

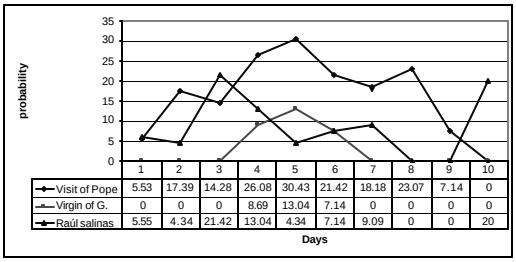
Abstract. A method for finding real-world associations between news topics (as distinguished from apparent associations caused by the constant size of the newspaper) is described. This is important for studying society interests.

Introduction.[*] Text mining is a new area of text processing that can be defined as discovery of interesting facts and new world knowledge from large text collections [2]. Its main tasks are analysis of trends, detection of deviations, and discovery of associations. In this paper, we focus on the analysis of a news collection. There are methods to detect new events in the news [1], to analyze news trends [4], and to separate good news from bad news [3]. The method we present allows analyzing a very common phenomenon in news: the influence of a peak news topic (important for a short time) over other news topics. This influence can show itself in two main ways: as a direct association (the peak topic induces other topics to emerge) and an inverse one (the peak topic causes temporal oblivion of another topic). We distinguish two different kinds of associations: observable ones (what we can see from the newspapers) and real world ones (between the events in the real world). The problem is that the observable association can be caused by the mere fact that all news reports must fill a constant size of the newspaper no matter what the total number of the events in the world is. This causes a *normalization effect* leading to apparent associations. In our model these two kinds of associations are proportional; thus the real ones can sometimes be inferred from the observable ones.

Finding real world associations. Given a peak news topic and the observable data surrounded, we can determine the absence or presence of a real-world association:

- Construct a probability distribution for each news topic over the time span corresponding to the peak topic. It expresses the probability of occurrence of this topic in each of the days of the period of interest.
- Detect the associations in the observed data set (if any). We use the correlation coefficient *r* for this.
- Estimate whether these observable associations are mainly due to a consequence of the normalization effect or there is a possible real-world association component.

[*] The work was done under partial support of CONACyT, REDII, and SNI, Mexico.



	1	2	3	4	5	6	7	8	9	10
Visit of Pope	5.53	17.39	14.28	26.08	30.43	21.42	18.18	23.07	7.14	0
Virgin of G.	0	0	0	8.69	13.04	7.14	0	0	0	0
Raúl salinas	5.55	4.34	21.42	13.04	4.34	7.14	9.09	0	0	20

Figure 1. Analysis of the peak topic *Visit of Pope*.

Basically, the latter estimation considers that: (1) any direct observable association has a high possibility for being a real-world one, since it goes against the normalization effect, and (2) any inverse observable association has a high possibility for being a real-world one if it does not disappear when the peak effect is eliminated.

Experimental results. Example: in the news from the Mexican newspaper "El Universal" for the ten days (January 20 to 29 of 1999) surrounding the *visit of Pope to Mexico City*, we found two real-world associations for this peak topic (Figure 1): a direct association with the topic *Virgin of Guadalupe* indicating that this topic probably emerged because of the influence of the peak topic, and an inverse association with the topic *Raúl Salinas* (Brother of the Mexican ex-president, sentenced on January 22). The latter association suggests that the topic *Raúl Salinas* went out of attention because of the influence of the visit of the Pope.

Conclusions. We have analyzed a phenomenon that is very frequent in real life situations – the influence of the peak news topics over other topics – and proposed a model in which the real-world associations can be inferred from the observable ones. Discovery of the real-world associations between news topics helps to interpret the social importance and influence of relevant but transitory topics and to define some parameters for better understanding of these news topics as well as our society on general.

References

1. Allan, J., Papka, R., and Lavrenko, V. (1998), Proc. Fo the 21st. ACM-SIGIR International Conference on Research and Development in Information Retrieval, Australia, 1998.
2. García-Menier E., "Un Sistema para la Clasificación de Notas Periodisticas", Memorias del Simposium Internacional de Computación CIC-98, México, D. F., 1998.
3. Hearst, M. (1999), "Untangling Text Data Mining", Proc. of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Marylnd, 1999.
4. Montes-y-Gómez, M., A. López-López, A. Gelbukh (1999a). "Text Mining as a Social Thermometer". In Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Stockholm, Sweden, 1999.