Classification Method for Shared Information on Twitter Without Text Data

Seigo Baba,
Fujio Toriumi,
Takeshi Sakaki
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan 113-8654
baba@crimson.q.t.u-tokyo.ac.jp,
tori@sys.t.u-tokyo.ac.jp,
sakaki@weblab.t.u-tokyo.ac.jp

Kosuke Shinoda,
Satoshi Kurihara
The University of Electro-Communications
1-5-1, Tyohugaoka, Tyohu-shi
Tokyo, Japan 182-0021
kosuke.shinoda@ni.is.uec.ac.jp,
kuri@is.uec.ac.jp

Kazuhiro Kazama
Wakayama University
930, Sakaedani,
Wakayama-shi
Wakayama, Japan 640-8441
kazama@sys.wakayama-u.ac.jp

Itsuki Noda
The National Institute of Advanced Industrial Science and Technology
1-1-, Umesono, Tsukuba-shi
Ibaraki, Japan 305-8568
i.noda@aist.go.jp

ABSTRACT

During a disaster, appropriate information must be collected. For example, victims and survivors require information about shelter locations and dangerous points or advice about protecting themselves. Rescuers need information about the details of volunteer activities and supplies, especially potential shortages. However, collecting such localized information is difficult from such mass media as TV and newspapers because they generally focus on information aimed at the general public. On the other hand, social media can attract more attention than mass media under these circumstances since they can provide such localized information. In this paper, we focus on Twitter, one of the most influential social media, as a source of local information. By assuming that users who retweet the same tweet are interested in the same topic, we can classify tweets that are required by users with similar interests based on retweets. Thus, we propose a novel tweet classification method that focuses on retweets without text mining. We linked tweets based on retweets to make a retweet network that connects similar tweets and extracted clusters that contain similar tweets from the constructed network by our clustering method. We also subjectively verified the validity of our proposed classification method. Our experiment verified that the ratio of the clusters whose tweets are mutually similar in the cluster to all clusters is very high and the similarities in each cluster are obvious. Finally, we calculated the linguistic similarities of the results to clarify our proposed method's features. Our method classified topic-similar tweets, even if they are not linguistically similar.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741726.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Analysis

Keywords

network, social media, data mining, clustering

1. INTRODUCTION

During such catastrophic natural disasters as earthquakes, tsunamis, and typhoons, victims and survivors must correctly and quickly collect information about shelters, dangerous areas, and safety advice immediately after disasters. Relief workers also need information about volunteers, relief goods, and providing food for evacuees. In other words, the required information changes based on the situations of those involved. However, such mass media sources as TV, newspapers, and radio offer general information instead of focusing on the more urgently needed local information.

Social media are attracting a great deal of attention since they can provide such localized information. In particular, many reports argue that Twitter, one of the most influential social media, is useful for sharing information during disasters. Mendoza et al. analyzed events related to the 2010 earthquake in Chile and characterized Twitter in the hours and days following it [1]. Miyabe et al. surveyed how people used Twitter after the 2011 Great East Japan Earthquake [2]. Sakaki et al. developed a novel earthquake reporting system that promptly notifies people of seismic activity by considering each Twitter user as a sensor [3].

In this paper, we also address Twitter as a source of local information. Previous works about extracting information from it focused on the text data of tweets. In other words, they were based on text mining. However, in some cases, text mining has difficulty extracting information. For example, it may be difficult to group tweets that are not linguistically similar by applying text mining. Consider the following tweets:

- Tweet 1: Don't go out after dark, especially during disasters.
- Tweet 2: Shut off the gas.
- Tweet 3: Remain calm.

These tweets can be categorized in the same cluster since they share the same topic: advice for victims immediately after a disaster. However, text mining cannot group them because of their poor linguistic similarities.

In this paper, we propose a novel tweet classification method that focuses on retweets without using text mining, based on the method proposed by Toriumi et al. [4]. By assuming that users who retweet a tweet are interested in it, we can classify the tweets that are required by users with similar interests based on retweets, even if they are not linguistically similar.

We subjectively confirmed whether our proposed classification is acceptable. Finally, we calculated the linguistic similarities of the results to clarify the features of the proposed method.

2. RELATED WORKS

Previous research exploited methods that extract information from Twitter. García et al. use the vector space model and Latent Dirichlet Allocation to obtain similar keywords[5]. Connor et al. found that consumer confidence and political opinions are contemporaneously correlated to sentiment word frequency in the texts of tweets [6]. Connor et al. clustered tweets using TweetMotif [7]. Tumasjan investigated whether Twitter is used as a forum for political deliberation and whether its online messages validly mirror offline political sentiment by analyzing the texts of tweets containing a reference to either a political party or a politician [8]. Rosa et al. researched automatic clustering and classified tweets into different categories by utilizing hashtags as indicators of topics [9].

Since these researches extracted information by focusing on the text data of tweets, they cannot group tweets with the same topics due to poor linguistic similarities. On the other hand, Toriumi et al. proposed a tweet classification method that focuses on retweets without using text mining [4]. However, their method requires a large amount of calculations and its validity has not been verified. Moreover, no evaluation of linguistic similarities was conducted. In this paper, we propose a novel tweet classification method based on the method proposed by Toriumi et al. and overcame such issues.

3. TWEET CLUSTERING

3.1 Data

In this paper, we use the log data of tweets written in Japanese that were posted and officially retweeted for 20 days from March 5 to 24, 2011. This period includes the Great Eastern Japan Earthquake that occurred on March 11, 2011. The log data contain 30,607,231 tweets. We selected tweets that were retweeted more than 100 times to focus on how the information was spread and shared. The number of such tweets is 34,860.

3.2 Constructing retweet networks

We constructed retweet networks based on the method proposed by Toriumi [4]. We used a bipartite graph [10] that consisted of tweets and the users who retweeted them to construct networks. When many users retweet both tweets A and B, they probably have a common interest in them and the topics are similar. In other words, two tweets whose similarity of retweet users is high may share a topic. Therefore, linking such tweets creates a retweet network that connects topic-similar tweets. The linking algorithm is based on co-citation algorithm which proposed by Small[11].

The similarity of retweet users between tweets t_i and t_j is defined as follows:

$$O_{ij} = \frac{|U_i| \cap |U_j|}{|U_i| \cup |U_j|}, \tag{1}$$

where U_i, U_j means users who retweeted t_i, t_j.

We applied the Jaccard coefficient [12] to the similarity. If O_{ij} exceeds a threshold value, t_i and t_j are connected. In this paper, we employed 0.05 as the threshold. It is presumed that the network structure depends on this value. It is one future work to clarify the relation between employing different thresholds and the network structure. We calculated the similarities for all the combinations of two tweets from the log data of tweets that were retweeted over 100 times. As the calculation results, the number of tweets with more than one link is 11,494 and the number of links is 30,363.

The constructed network is shown in Fig. 1. Each node represents a tweet, and each edge represents a link between tweets whose degree of similarity of retweet users is over the threshold. The size of the communities is different. Communities with a few nodes are shown in the lower part, and communities with many nodes are shown in the upper part.

3.3 Extracting similar tweets from retweet networks

Among the communities obtained by constructing a retweet network, some have many nodes. Such communities are located in the upper left of Fig. 1.We assume that such large communities have various topics.

We applied our clustering method to the entire area of the retweet network to extract clusters that contain similar tweets. Our clustering method is expanded from Newman's method [13], which is based on modularity [13]. Modularity is a property of a network and a specific proposed division of it into communities. The higher it is, the better the division is, in the sense that the number of edges within communities is more than we expect by chance and the number of edges between them is less. Modularity is defined as follows:

$$Q = \frac{1}{2M} \sum_{c=1}^{N_{CM}} [\sum_{i,j=1;n_i,n_j \in CM_c}^{N} (A_{ij} - \frac{k_i k_j}{2M})], \tag{2}$$

where $N, M, A, n_i, k_i, N_{CM}$,and CM_c respectively denote the total number of nodes, the total number of edges, the adjacency matrix of the network, node i, its degrees, the total number of clusters, and cluster c.

The algorithm of the Newman method is defined as follows:

1. Generate N clusters so that each node belongs to one cluster.
2. Loop action begins below.
   (a) Select cluster i.
   (b) Select cluster j among the adjacents of cluster i.

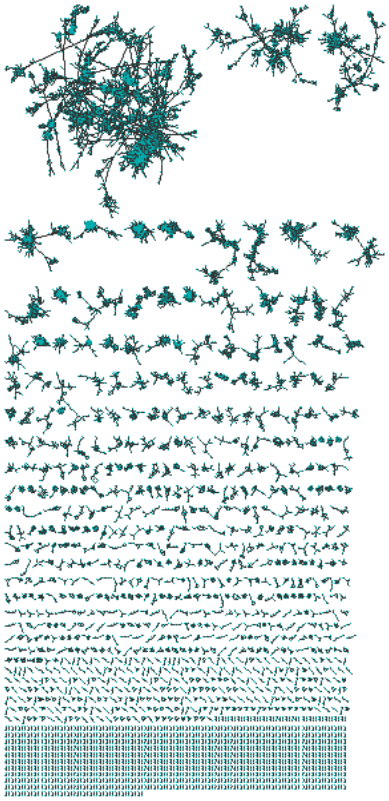

Figure 1: Retweet networks

(c) Calculate ΔQ_{ij} if cluster i is combined with j.

(d) Repeat steps (a), (b), and (c) for all the combinations of two clusters.

(e) Select the largest ΔQ_{ij} and combine clusters $i and j$.

(f) Calculate Q

(g) Repeat these steps until only one community remains.

3. Output the community structure of the largest Q.

In this paper, we applied the clustering method by adapting Newman's method to reduce the amount of calculations. The method's algorithm is defined as follows:

1. Generate N clusters so that each node belongs to one cluster.
2. Loop action begins below.
   (a) Select cluster i.
   (b) Select cluster j among the adjacents of cluster i.
   (c) Calculate ΔQ_{ij} if cluster i is combined with j.
   (d) Repeat steps (a), (b), and (c) for all the combinations of two clusters.
   (e) Select the largest ΔQ_{ij}:
      i. for the largest $\Delta Q_{ij} > 0$, combine clusters $i and j$ and repeat steps (a) ~ (e).
      ii. for the largest $\Delta Q_{ij} \leq 0$, quit the loop action.
3. Output the community structure.

Table 1: Example of a question

Statement tweet	The site gives information about the distance between your place and the Fukushima No.1 nuclear power plant and rolling blackouts http://gigaz.in/KJm8j
Tweet A	Twitter is a source of information
Tweet B	Yahoo! Map shows the area of the rolling blackouts of the Tokyo Power Company http://gigaz.in/KHzL4

This method reduces the amount of calculations more than the Newman method.

We found 2,001 clusters after applying the clustering method.

4. SUBJECTIVE EXPERIMENTS

The ratio of clusters whose nodes are mutually similar in the cluster to all clusters is not clear. In the same way, whether the similarities of the nodes in each cluster are obvious has not been verified either. In other words, the proposed method's validity has not been clarified. Thus, we conducted a subjective experiment to do so. In this section, we first describe how we conducted the experiment in Section 4.1 and discuss the results in Section 4.2.

4.1 Details of subjective experiment

The following are the details of our experiment. Its question consists of three tweets: one statement tweet and two choice tweets. The examinees selected a choice tweet whose topic most closely resembles the statement tweet from the following choice tweets:

- Inner tweet: belongs to the cluster to which the statement tweet belongs.
- Outer tweet: belongs to the cluster to which the statement tweet does not belong.

If an examinee selects the inner tweet, his judgment corresponds with the results of the proposed method. An example of the questions is shown in Table 1. The tweets were written originally in Japanese, but the samples in this paper were translated into English.

Fourteen people participated in this experiment as examinees, and each question was solved by seven examinees. If more than four examinees selected the inner tweet of each question, the question was labeled as correct. Each examinee worked questions randomly to exclude the impact of the order on the result. We randomly selected 100 questions from all the tweets, and each examinee solved 60 of them. The first ten questions of the 60 were not added to the result to avoid any influence on the examinees by selection standard during the early solving stage.

4.2 Subjective experiment results

From the experiment results, 89% of all the questions were correct in the totalization result. In some cases, the three tweets (the statement tweet and two choice tweets) were similar mutually among the incorrect tweets. Table 2 shows an example of such cases. The topics of the three tweets were advice for victims immediately after the disaster. In our proposed method's results, some clusters have identical topics. That is, these questions were only incorrect in the assessment of this experiment, but actually the statement tweets and their inner tweets were similar. Solving that problem is future work.

The relation between the rate of questions and the number of corresponding examinees is shown in Fig. 2. When four or five

Table 2: Example of a question where topics of three tweets are similar

Statement tweet	[Please spread] If you are not able to move, please use the tag [#j_j_helpme] and post tweets with your GPS information, if you can.
Tweet A	[Please spread] Women should not go out after dark. During disasters, sexual predators may pretend to be volunteers.
Tweet B	Please calm down, wear thermal clothing, and pack money, valuables, food, water, and a mobile phone with a battery charger.

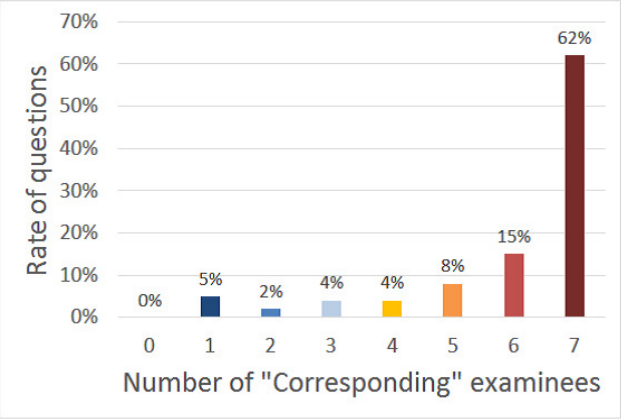

Figure 2: Relation between rate of questions and corresponding examinees

examinees selected the inner tweet, the similarity of the nodes in the cluster is not obvious. However, in 77% of the questions, more than six examinees selected the inner tweet. From this result, we conclude that the similarities of the nodes in each cluster are obvious.

Thus, we confirmed the validity of the proposed method where the rate of the clusters whose nodes are mutually similar in the cluster to all clusters is very high and the similarities of the nodes in each cluster are obvious.

5. CALCULATION OF LINGUISTIC SIMILARITIES OF CLUSTERS

Some clusters have nodes with little linguistic similarity. Such samples of tweets are shown in Table 3. The cluster in example 1 contains tweets that have the same topics about shelter. However, they share little linguistic similarity. The cluster in example 2 also contains tweets about advice for victims and their linguistic similarity is also poor. These tweets were grouped in the same cluster since they drew attention from users who retweeted them under similar situations despite their low linguistic similarities. Thus, in this section we calculated the linguistic similarities of the results to clarify the features of the proposed method by applying a vector space model [14] based on TF-IDF [15].

Table 3: Clusters of tweets

Example 1: cluster that groups tweets about shelter

The Oura cafeteria on the Ueno Campus of the Tokyo University of the Arts is open. You can spend the night there.
[a quick report] Okumakodo is open! It looks like it has some blankets http://twitpic.com/48f6y2
Are you all right? [The Tokyo Bunka Kaikan just opened. It's getting dark and cold, so if you are around Ueno Station, please go there.]

Example 2: cluster that groups tweets about advice for victims

If you are evacuating with a baby, wrap the baby in a blanket and carry it in a tote bag. No baby buggies! #jishin
[Please spread] If you use Twitter by mobile phone, turn off your icons to conserve battery life.

5.1 TF-IDF

Essentially, TF-IDF determines the relative frequency of words in a specific document compared to its inverse proportion over the entire document corpus. This calculation intuitively determines how relevant a given word is in a particular document. Words that are common in a single or a small group of documents tend to have higher TF-IDF numbers than such common words as articles and prepositions [15]. TF-IDF consists of Term Frequency (TF), which is a term's frequency in a document, and Inverse Document Frequency (IDF), which is the inverse of the frequency of a document that contains the term in all documents. The TF and IDF of term t in document d are defined as follows:

$$tf(t,d) = \frac{n_{t,d}}{\sum_{i \in d} n_{i,d}} \tag{3}$$

$$idf(t) = \log_2 \frac{N}{df(t)} + 1, \tag{4}$$

where $n_{t,d}$, N, and $df(t)$ respectively denote the number of occurrences of term t in document d, the number of total documents, and the number of documents that include term t.

5.2 Vector space model

A vector space model calculates the linguistic similarity between two documents. In this paper, we regard a tweet as a document and calculate the linguistic similarity of two tweets. The following are the method's details:

1. Do morphological analysis on all tweets.
2. Generate a feature vector of N dimension v_i of tweet T_i, where N is the number of morphemes of all the tweets:
   - $v_i = (w(t_{1,i}), w(t_{2,i}), , , , w(t_{N,i}))$
   - $w(t_{k,i})$ is the TF-IDF of term $t_{k,i}$ in tweet T_i if T_i does not include $t_{k,i}$, $w(t_{k,i}) = 0$.
3. Calculate:
$$\cos\theta = \frac{v_i \cdot v_j}{|v_i||v_j|} \tag{5}$$
as the linguistic similarity between T_i and T_j.

5.3 Evaluation results of linguistic similarity

We applied the vector space model to the results of the proposed method to evaluate the linguistic similarities of the results. The

Table 4: Samples of clusters

Example 1: clusters whose linguistic similarity is 0.044

This tweet was posted by a volunteer center. Yesterday, more than 1000 people read it and learned about dangerous areas and shortages. What should we do? http://t.co/4JpWlXt #jishin
RT [please spread] If you want non-allergic milk or alpha rice, please call 0524855208 or mail info@alle-net.com. The building where allergy treatments take place in Nagoya will send them!
RT [please spread] Check that your car has a jack for changing tires. They are useful for rescuing victims from rubble. #jishin #jisin

Example 2: clusters whose linguistic similarity is 0.0108

If children are shaking or suffering from fear, hug and comfort them.
I've experienced two big earthquakes. I spent a few nights in a car and saw many senior citizens who seemed to be suffering from economy class syndrome from remaining in the same posture for a long time. If you have to spend too much time in a car or a cramped shelter, don't forget to stretch your legs.

Example 3: clusters whose linguistic similarity is 0.0052

RT [Summarize the information on TimeLine]◯open the door◯cook some rice◯place baggages in an entrance◯buy water, snacks and a towel◯blankets◯a flashlight◯a thin plastic film made of saran◯wear shoes◯store water in a bath◯charge a mobile phone◯switch off an ampere breaker◯close the gas tap◯refrain from use of the cellular phone◯pay attention to bits of glass◯calm down[add more, please]
My friend who survived the Great Hanshin Earthquake evacuated his house in pajamas. So tonight, sleep in clothes just case you have to leave quickly.

number of morphemes of all the tweets was 50,731 after morphological analysis on the log data used in this research: in other words, the data of 34,860 tweets described in Section 3.1. Then we generated the feature vectors of the 50,731 dimensions of each tweet and calculated the linguistic similarities of two tweets for all the combinations of all 34,860 tweets, including the linked and unlinked combinations to make reference values. Their average was 0.0156 and their standard deviation was 0.0218. When the similarity between two tweets is under the sum of two values (0.0374), their linguistic similarity is random at most.

We calculated the linguistic similarities of the links connected in Section 3.2. 31.0% of all the links are under 0.0374. This means that some of the links connected by this proposed method are the edges of two tweets, which are difficult to link by text mining.

We also calculated the linguistic similarities in each cluster. The linguistic similarity in a cluster is defined as the average of the tweets for all the combinations of the nodes that belong to the cluster; for example, that of the cluster that contains four nodes is the average of ${}_4C_2$ combinations). 19.1% of all clusters are under 0.0374. This means that some of the clusters obtained by this proposed method group tweets whose linguistic similarities are as low as randomly selected tweets. In other words, the proposed method identifies clusters with low linguistic similarities, but high similarities from the viewpoint of information.

Table 4 shows samples of the tweets of clusters whose nodes only share slight linguistic similarity. As the table shows, even though these samples are not linguistically similar, they do share similar topics that attract the same interests. The topic of example 1 is about aid to volunteers and getting supplies, example 2 is offering advice about life in shelters, and example 3 is providing advice for victims. This confirms that our proposed method can classify topic-similar tweet users who share the same position needs, even if they are not linguistically similar.

6. CONCLUSIONS

In this paper, we proposed a novel method of the classification of tweets by focusing on retweets without using text mining.

We conducted a subjective experiment to confirm the validity of the proposed classification method. The ratio of clusters whose nodes are similar to each other in the cluster to all clusters is 89% and the similarities of the nodes in each cluster are obvious.

We also calculated the linguistic similarities of the results and applied a vector space model based on TF-IDF which determines the relative frequency of words in a specific document compared to its inverse proportion over the entire document corpus. We confirmed that our method can classify topic-similar tweet users who have the same situation needs, even if they are not linguistically similar.

We employed 0.05 as the threshold for the similarities of retweet users between two tweets. It is presumed that the network structure depends on this value. It is one future work to clarify the relation between employing different thresholds and the network structure.

Some clusters should be grouped as layered structures. For example, clusters about advice for victims immediately after an earthquake and those about shelters may be in a cluster that groups information for victims. Thus, future work will investigate a clustering method that can extract such a layered structure.

Moreover, in some cases, the questions in our subjective experiment are incorrect, but these statement tweets and inner tweets have similar information. Future work will also conduct another subjective experiment and consider such cases.

Finally, this proposed method will apply a system that provides similar information for disaster situations. Under such situations, information must be provided quickly. Reducing the amount of calculations is a critical future work.

7. ACKNOWLEDGMENTS

We thank Genta Kaneyama (Cookpad Inc.) for assistance in collecting data from Twitter. We also thank the people who supported this research.

8. REFERENCES

[1] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we RT? In Proceedings of the First Workshop on Social Media Analytics -SOMA 10, pages 71-79. ACM Press, July 2010.

[2] M. Miyabe, E. Aramaki, and A. Miura. Use trend analysis of twitter after the great east japan earthquake. In Proceedings of SIG-DPS/GN 2011-DPS-148/2011-GN-81/2011-EIP-53, 2011.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, WWW ′ 10, pages 851-860. ACM, 2010.

[4] Fujio Toriumi, Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, and Itsuki Noda. Information Sharing on Twitter During the 2011 Catastrophic Earthquake. 2nd International Workshop on Social Web for Disaster Management (swdm2013) WWW ′2013 Companion Publication pp.1025-1028

[5] García-Silva, A., Kang, J. H., Lerman, K., and Corcho, O. Characterising emergent semantics in twitter lists. In The Semantic Web: Research and Applications (pp. 530-544). Springer Berlin Heidelberg, 2012.

[6] B. O. Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series. Most, pages 122-129, 2010.

[7] B. O. Connor, Krieger, M. , Ahn, D. Tweetmotif: Exploratory search and topic summarization for twitter, Proceedings of ICWSM, pp. 2-3 (2010)

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment. Word Journal Of The International Linguistic Association, pages 178-185, 2010.

[9] Rosa, K. D. , Shah, R. , Lin, B. , Gershman, A. , and Frederking, R. Topical clustering of tweets, in Pro-ceedings of SIGIR Workshop on Social Web Search and Mining (2011)

[10] S. Wasserman and K. Faust. Social network analysis: Methods and applications. In Structural Analysis in the Social Sciences, volume 8, pages 299-302. Cambridge University Press, 1994.

[11] Small HENRY. Co?citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265-269, 1973.

[12] W. B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall PTR, 1992.

[13] Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks, *Physical review E*, Vol. 70, No. 6, p. 066111 (2004)

[14] G.Salton, A.Wong, C.S.Yang. A Vector Space Model for Automatic Indexing(1975)

[15] Juan Ramos: Using TF-IDF to Determine Word Relevance in Document Queries(2003)