http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (2): 236-240, 2009
ISSN 1812-5638

The Algorithm of Short Message Hot Topic Detection Based on Feature Association

[1,2]Qindong Sun, [1]Qian Wang and [1]Hongli Qiao
[1]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]Network Computing and Security Key Lab of Shaanxi, Xi'an University of Technology,
Xi'an 710048, China

Abstract: Aiming at the mobile short message (SMS) hot topic extraction, the text features and statistical regularity of SMS are analyzed in this study. The formal description of SMS hot topic is given and an algorithm of SMS hot topic extraction based on feature association analysis is proposed. According to the proposed algorithm, feature words of SMS can be clustered into different word bags by calculating the association degree of these feature words and the hot topic can be identified by means of word bags matching. Experiments results show that the proposed algorithm can detect the hot topic in the SMS messages effectively, which is useful to the analysis of SMS popular sentiments.

Key words: Short message, hot topic, features association, popular sentiments

INTRODUCTION

As a kind of new media, mobile short message is attracting more and more users and its influence is enlarging day by day. People use short messages to talk about their life and social hotspot. So short message stream contains public opinions. Earthquake prediction and poisonous bananas are examples of short message hot topics. The effective topic detection and tracking (TDT) to short message stream can help people to supervise such information.

TDT is a hotspot in the research of information processing area. Many scholars have studied relative technologies on Internet information and newspapers. Walls *et al.* (1999) has employed probabilistic model to detect topics of broadcast news and four alternated versions of this model are also discussed. Victor *et al.* (2002) has applied the technology of correlation modeling to TDT correlative detection tasks and has got distinct improvement. Mouri and Hiroyuki (2004) has talked about the technology of finding new topics in hidden websites with biased inquiry probes. Hamamoto *et al.* (2005) has used technologies of SVD, clustering and independent component analysis to extract features in text stream and then done the topic detection experiments followed by comparison and analysis. Allan *et al.* (2005) has talked about the shortcomings of open assistant optimization TDT system under practical circumstances and brought forward a resolution of dirty clustering problem. Chen *et al.* (2007) has provided a hot topic extracting algorithm within a given period time in independent text sets to deal with real-time problem of Internet information tracking. Yun *et al.* (2008) has applied respective method of latent semantic analysis and probability based latent semantic analysis to fetch BLOG hot topics. Zhou *et al.* (2007) has been provided a hot word relation based Internet hot topic detection algorithm to answer the requirement of network public opinion analysis. According to the relationship between BLOG comments and BLOG topics, Shi *et al.* (2008) has researched the technology of BLOG hot topic detection and public opinion supervising.

Researches on short message TDT are still deficient now. Wu *et al.* (2007) has studied physical characters in short message complicated network. Compared with Internet and newspaper information, short message texts and its topic spreading have particular characters. So the researches above cannot be applied to short message TDT directly.

In this study, a short message hot topic detection algorithm is proposed, which based on content feature association via analyzing the short message text's characters and related statistical properties. It can also be helpful in the further research of short message public opinion analysis and in the design of new information service to meet mobile users' increasing needs.

FORMAL EXPRESSION OF SMS HOT TOPICS

The definition of short message hot topic and its formal description are bases of short message TDT. Short message hot topic has commonness with that in other

Corresponding Author: Qindong Sun, School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

media but it also has some independent properties coming from the structure of short message. So, a proper SMS hot topic definition considering its idiosyncrasies is given, which is different from that of Internet and newspaper media.

Definition

Short Message Hot Topic (SMHT): The short message stream in a given period contains some content focuses whose content is homogeneous and the number of participants is over a given threshold. Each of them can be called a Short Message Hot Topic (SMHT).

All short messages are text strings of natural language sentences and each sentence can be divided into a group of notional words together with a group of empty words. Short messages with the similar content may not be identical, but they should contain elements which are the same or expressing the same meaning. So a SMHT can be expressed by a group of feature words generated from short messages of this topic. These words must be relevant in content. The collection of those words is called as a Word Bag (WB).

$$WB = \{WF_1, WF_2, WF_3, ...\} \quad (1)$$

where, WF_i is the number i feature word. Each short message topic has a WB of feature words and it also needs other factors as follows: CSM, a non-empty short message set as topic participants; N_{csm}, the count of participant short messages as the measure of topic hot degree; T, the time span of short message stream is also an independent factor of the topic. So a short message topic could be formatted as follows:

$$SMT(TC) = \{CSM, N_{csm}, T, WB\} \quad (2)$$

where, TC is the type of topic. From discussions above we know SMHT is the SMT with concentrated content and a good many participants. We can set a threshold N_{Gate} to denote topic participants, thus SMHT can be formatted as:

$$SMHT(C) = (CSM, N_{CSM}, T, WB) \ \ N_{CSM} >= N_{Gate} \quad (3)$$

ASSOCIATION DEGREE COMPUTING OF SMHT FEATURE WORDS

The content of short messages is composed of text, whose words has specific meaning. Generally, short messages are short and information concentrated, so its each word is more importance in expression. Short messages can be delegated by its feature words. Thus a counterpoint relationship can be established among WB, CSM and TC in the same topic:

$$CSM \Leftrightarrow WB \Leftrightarrow TC \quad (4)$$

The short messages content are complex and diversified, so it is hard to classify the short message stream with supervised machine learning algorithms. In this paper, feature words are viewed as the counterpoint of SMT and the hot topic in message stream is recognized on basis of the feature word bag of each topic.

As far as unsupervised machine learning is concerned, it is infeasible to get the feature words bag correspond to each topic directly. However by means of word frequency stat, the collection of all feature words in the whole short message corpus could be found out, which is labeled as CFW. Then CFW could be divided into word bags according to content association degree:

$$CFW = \cup WB \left(i \neq j, WB_i \cap WB_j = \varnothing\right) \quad (5)$$

Each WB corresponds to a short message topic. By pattern matching of word bags on short message corpus, the participations of each topic can be found out. The division of CFW is based on the association degree of separate feature words. In previous research, the association degree is calculated as following:

$$Leverage\ (X, Y) = P(X,Y) - P(X)P(Y) \quad (6)$$

By given an appropriate Leverage threshold, features of the same topic can be identified. But there are still some problems if the result of Eq. 6 is directly used as the measurement of association degree. The feature word set is only a small subset of all the words and there is great difference between their Document Frequency (DF) values. Therefore the difference of their P(x, y) values is great too, which leads to the leverage (x, y) value could not reflect the association relationship of words. In this study, the original leverage formula is improved. The relative leverage value is used as the measurement.

Theorem: Given a short message corpus, if Leverage (x, y) is employed to represent the association of two feature words X and Y, whose DF values are known, there will be a maximum value of two feature words' association degree.

Prove: Let the size of corpus is U, features' document frequency is labeled as DF and the set of all short messages containing feature words is SM, then:

$$\exists\ x, y,\quad DF(x) > DF(y) \geq 1$$

$$\because\ P(x) = {DF(x)}/{U},\quad P(y) = {DF(y)}/{U}$$

$$\therefore\ P(x) > P(y)$$

$$Leverage(x,y) = P(x,y) - P(x) \times P(y)$$

$$\because\ DF(x, y) \leq DF(x),\ DF(x,y) \leq DF(y),\ P(x,y) = {DF(x,y)}/{U}$$

$$\therefore\ P(x,y) \leq P(x),\ P(x,y) \leq P(y)$$

$$\because\ P(x) > P(y)$$

$$\therefore\ 0 \leq P(x,y) \leq P(y)$$

$$\therefore\ Leverage(x,y) \leq P(y) - P(x) \times P(y)$$

It is obvious that max Leverage (x, y) is obtained by the smaller probability values of x and y minus the product of these two probability values. The case of maximum Leverage (x, y) is got when SM (x) ⊇ SM (y) and association between feature X and feature Y is the closest. This max value is labeled as Max (x, y). Then the association degree of x and y could be calculated with the following formula:

$$RL\ (x, y) = \frac{Leverage\ (x, y)}{Max\ (x,y)} \qquad (7)$$

By used formula 7, the impact of DF difference to association degree computation is wakened.

THE DETECTING ALGORITHM OF HOT TOPIC BASED ON FEATURES ASSOCIATION

Section above has talked about the method of features' association computation. After got association value (RL) of all feature pairs, further processions are needed to find hot topics from short message stream. Owing to the specialty of short message information, more pretreatments are taken on the short message corpus and its words.

Firstly, some short messages are too short to express an independent meaning and could not be used to detect short message hot topics, so they will be eliminated from the corpus. Secondly, as stated above short messages need to be split into single words. Some words like a, an, the, but, which have no semantic significance are filtered. Then computing of frequency is done on left words.

Algorithm description

Input data: The corpus of short messages labeled as CSM and the corpus feature words set CFW.

Expected output: Feature WB of each SMT and HSM, the set of participants of each SMT.

Step 1: Get CFW through the preprocessing on CSM. Then, the CFW composes a vector S with a dimension N;

Step 2: Buildup a N*N numerical matrix M. The number i row of this matrix express the association degree between number i feature and other features. The number j line stores the association degree between other features and number j feature word. So M_{ij} express the association between number i feature and number j feature. The initial value of matrix elements is 0.

Step 3: Read one short message SM from CSM and match it with feature vector S. Those features sets of S whose match is success can be labeled as S_{match}, $S_{match} = \{WF_1, WF_2, \ldots WF_n \mid WF_1 \in CFW, WF_j \in SM\}$. Select all kinds of feature pairs in S_{match} and record their index in S and $\forall WF_i, WF_j \in S_{match}, i \neq j, M_{i,j} + +, M_{ij}$ will be added by 1. When each SM of CSM is matched, the matrix M will store the association count of each feature pairs.

Step 4: Traverse the matrix, for each non-zero element, Max (WF_i, WF_j) is computed with formula 6 and then compute RL (WF_i, WF_j) with formula 7.

Step 5: Set association threshold on RL value and traverse each row of the matrix, merge all features which are greater than the threshold into the same feature subset of WF standard element of this row and will get an independent topic word bag WB. Labeling each feature of this WB to indicate that this feature has been a member of certain SMT and it will never attend other topics any more.

Step 6: Subsets of features is the result of a division of S, i.e., CFW = ∪WB. If number of features in WB is less then a given threshold N_{WBGate}, this WB will be invalid. So all useful feature word bag meets the condition: WB ($N_{WB} \geq N_{WBGate}$).

Step 7: Match every short message in CSM with word bag of SMT to find out which topic it belongs to. Then short messages related with this SMT (HSM) will be found. And the property of matching result is CSM ⊇ ∪HSM.

Step 8: Set the threshold of SMHT as compares it with number of each topic's short message, then effective SMHTs are selected out i.e., HSM ($N_{SCSM} > N_{SMGate}$).

Step 9: Label each effective short message hot topic, SMHT = (HSM, N_{SCSM}, T, WB, TC).

Additionally, the front 7 steps are processed with machine completely and the last two steps require some manual job.

Table 1: Statistics of experimental result

SMTH	No.	Proportion (%)
Valid	13	26.5
Invalid	1	2.0
Non	34	69.4

Table 2: List of valid short message hot topic

No. of participants	Content of SMHT	Typical features in WB
169	Holiday wishes	Happy spring festival, wish, happy new year
312	Affective interaction	Glad, missing you, sad, honey, friendship
275	OA system information	Communication, YAN'AN, Sales, Series, operation, transmit, HANZHONG
206	Joke about animal	Turtles, mouse, super, rabbit, fool
263	Fetion register	Visitor, good friends, online, nickname, loneliness, agree, age
71	Life message	Drink, son, concomitance, always, early, specially
134	SP service customization	Account, securities, transactions, any time, the cost of information, call
131	Campus communications	Snow, schools, school, brother, to accompany me in the morning, poor, Internet
70	Joke	TV, beautiful women, calls, watching TV, dream
104	Recommended music ringtones	Song, wonderful, ring tones, enjoy, Phoenix, moon
92	Oil reserve inform	Enterprise, security, oil, obligations, supply, stabilize, the proposed oil
222	SMS fortune-telling	Ame, comment, suggested that the note
133	Insurance inform	Pay, for life, in order to avoid, coming

RESULTS AND DISCUSSION

The experimental corpus comes from one mobile communications company and the total number of short messages is 165411.

The short messages whose length is less than 20 had been eliminated. The amount of selected features is 1000 and the associate threshold is set as 0.8 and the threshold of participants is set as 50.

The result of this algorithm is shown in Table 1. Forty-nine topic are extracted out. The rate of accuracy of SMHT detection is 92.9%. Most hot topics are valid and have concentrated theme, as shown in Table 2.

The results show that short message texts are rich in content and various in topic. If the number of topic participants is great, the topic content will be clear and remarkable, otherwise it will be hard to extract the topic from short message corpus. From Table 2 we also can notice that although short messages in CSM is as much as more than 160 thousand, there are only hundreds of participants for each hot topic. SMHT short messages take only a small part of the whole corpus. The widespread of message content makes obvious difficulty in short message classification and SMT detection.

CONCLUSION

To research the problem of short message hot topic detection and extraction, the character of short message text and statistical properties of short message words are analyzed. A formal description of short message hot topic is provided and a detecting algorithm of short message hot topic based on features correlation is put forward. In this algorithm, the association degree of different feature pairs is calculated to split feature words set into hot topic word bags and participants of each topic are ascertained by matching feature words bag with messages corpus. Whether a topic is hot or not is determined by the number of its participants. Experimental result shows that the proposed algorithm can effectively find hot topics in the short message stream.

The further research will primarily focus on how to extract features in short message corpus more effectively.

ACKNOWLEDGMENTS

The research work in this article is partially supported by the National Natural Science Foundation of China (No. 60802056); the Natural Science Foundation of Shaanxi Province (No. 2007F13), the Industrialization Project of Shaanxi Province Education Office (No. 08JC09).

REFERENCES

Allan, J., S. Harding, D. Fisher, A. Bolivar and S. Guzman-Lara, 2005. Taking topic detection from evaluation to practice. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Jan. 03-06, IEEE Computer Society Press, pp: 101-101.

Chen, K.Y., L. Luesukprasert and S.C.T. Chou, 2007. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE Trans. Knowledge Data Eng., 19: 1016-1025.

Hamamoto, M., H. Kitagawa, J.Y. Pan and C. Faloutsos, 2005. A comparative study of feature vector-based topic detection schemes for text streams. Proceeding of International Workshop on Challenges in Web Information Retrieval and Integration, Apr. 8-9, IEEE Computer Society Press, pp: 122-127.

Mouri, T. and K. Hiroyuki, 2004. Extracting new topic contents from hidden web sites. Proceeding of International Conference on Information Technology: Coding Computing, Apr. 5-7, IEEE Computer Society Press, pp: 314-319.

Shi, D.M., H.F. Lin and J. Zhao, 2008. Blog information extraction based on template. Comput. Eng. Appli., 44: 156-158.

Victor, L., A. James, D. Edward, V. Pollar and S. Thomas, 2002. Relevance models for topic detection and tracking. Proceeding of the Human Language Technology Conference, Mar. 24-27, San Diego, USA., pp: 115-121.

Walls, F., H. Jin, S. Sista and R. Schwartz, 1999. Probabilistic models for topic detection and tracking. IEEE Piscataway, 1: 521-524.

Wu, Y., J.H. Xiao, Y.Z. Wu and J.Z. Yang, 2007. Research on the growing process of short message networks. Acta Phys. Sinica, 56: 2037-2041.

Yun, C., S. Flora Tsai and C. Kap Luk, 2008. Machine learning techniques for business blog search and mining. Expert Syst. Appli., 35: 581-590.

Zhou, Y., S. Qindong and G. Xiaohong, 2007. Internet popular topics extraction of traffic content words correlation. J. Xi Jiaotong Univ., 41: 1142-1145.