Research and Apply Language Rhythm in Topic Tracking

Fan Chen

Information Science & Technology Department, Tianjin University of Finance and Economics, Tianjin 300200, China

Keywords: Topic Tracking, Language Rhythm, Rhythm Characteristics, State Switch.

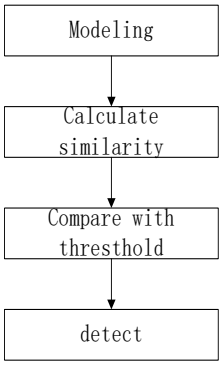
Abstract. Language Rhythm is an important characteristic in language. Research and analysis the Language Rhythm, those four kinds of language rhythms prompted: language nature rhythm, language grammar rhythm, language logic rhythm and language emotion rhythm. Each language rhythms reflect the language characteristic. How to get the language rhythms and their characteristics is expounded. The simulation results show that it is significant to apply in topic tracking.

1. Introduction

Today, there is massive information on Internet. People are not worry about how to find the information, but how to located they interested. And how to locate the interested information for people quickly and accurately has become the hot research subject. Topic Tracking (TT) is one task of Topic Detection and Tracking (TDT) [1], which focuses on how to find some other reports, happened on different time about the same topic. But the all kinds of discussion about the same topic are not occurred in the same site or same time, they are maybe in anywhere in Internet at any time. So it is an important task that how to find the isolated and related reports about the same topic. The definition of topic in TDT is one core event and the other event s which are associated with it [2]. While each event is always caused by some reasons and conditions and happened on some time at some locations with someone or something. Then, the topic can be regarded as a collection of a number of related reports on the same event [3]. In this paper, a new method is proposed in topic tracking witch is different with the other ones. A new feature of language is used in TT, and it is Language Rhythm. Language Rhythm is the nature character of language, everyone cannot express himself or herself clear when talking without Language Rhythm and each article cannot be understand without Language Rhythm. So how to get the Language Rhythm and how to use it in TT are the main event in this paper.

2. Topic Tracking System Process

Usually, the Topic Tracking system process is contained four steps: the first is creating the models of the topic which is waiting for is traced and the reports which are ready to detect, the second is calculated the similarity of the two models, the third is comparing the result with the threshold and the last is when the result reaches the threshold, then it can be detected that the report is belong to the topic [4]. As shown in figure 1 below:



Figue 1 topic tracking system process

© 2015. The authors - Published by Atlantis Press

In Topic Tracking, it is important and difficult that how to find the correlation between the reports and topic. A useful modelling method can express the topic and report effectively and lots of great method can be used in calculating the similarity, and then to find that the reports are about the topic or not. Vector Space Model (VSM) [5] is a good model method; the similarity can be calculated by cosine similarity and by Hellinger distance [6]. Blei proposed the Topic Model [7] in mining the distribution of the topics in texts and Martijn and etc. used the language model [8] in Topic Tracking. But all these methods are very useful in detecting the similarity of two or more reports and documents and etc. But the reports about one topic maybe very different in text ,so it is not a good method in topic tracking only comparing the similarity of the reports in text. Then, it is the difficult and important that how to break the limitation of content similarity in the topic developing. As a result, it attracts more research that how to find the reports which are talking about the same topic within a certain time limit.

3. Language Rhythm

Language Rhythm is a nature feature and phenomenon in human language. Each document has the unique rhythm, but there are still some similarity in the reports and documents of the same topic. It can be used in the task of Topic Tracking. As time going on, the content of report about one topic maybe changed, but the Language Rhythm maintains some consistency usually. So Language Rhythm can be used in Topic Tacking.

Language is on the order of time whether from the speech carrier or the text carrier, so it is the organization of time. And this order is very useful in express the sematic and emotion in the reports or documents. Language Rhythm can indicate this order. It is the organization feature of language in timeline, and it can mark the language order with rhythm characters from physical, logical and emotional. It is the key that how to find the Language Rhythm and which Rhythms are useful in topic tracking.

3.1 Classification of language rhythm

There are four language rhythms, nature rhythm, gramma rhythm, logic rhythm and emotion rhythm.

Nature rhythm is the basic of language rhythm. It reflects some physical features such as breathe rhythm and some hidden logic features, and even emotions. So it is the most important and common one in language rhythm. Which element in language can mark it? These elements must be existed in each report and document, but their sequences are must be different. Then punctuations can mark the nature rhythm successful, because that they can mark different pauses as the necessary for expression and physician.

Gramma rhythm are marked by the auxiliary words which maybe is no useful in sematic but is very important in constructer .They can indicate the relationship and attribute of the contents linked by them. There are lots of auxiliary words in each report and document, but the amount of them is so limited far less than the other words. So they can be used in marking Language Rhythm, too.

Logic rhythm also is an important rhythm in language. When persons want to expresses something they must be in their own logic. The report and document cannot be understood without logic .So logic characters can be used to mark the logic rhythm. Some special words can be used, such as “because” ,”and” ,”the first” and etc.

Emotion rhythm is the high level rhythm in language. Its existence is depended on necessary, so it often absence in some occasion. But when the people are faced one topic, they probably have the same kind emotion, and then reflect in the words .Then, emotion rhythm can indicate the same emotion of different person. It is very useful in topic tracking. Some emotion words can be used to mark emotion rhythm.

3.2 Obtain Language Rhythm

Language Rhythm is ordered in timeline, and then it can be calculated as a first-order Markov process. In the language rhythm array, each rhythm is in its position randomly and correctly. The positions of marks maybe are usefulness in sematic but they reflect the other characters in language, reports and documents. As a result, analyzing and calculating the state transition probability of the

marks 'positions can be used as the method to obtain the language rhythm. The k mark appears is only associated with the k-1one, as shown in formula 1 below.

$$P(Mark(RU_k)) = P(Mark(RU_k) = mark_j \mid Mark(RU_k) = mark_i, Mark(RU_{k-2}) = mark_k, \cdots) \quad (1)$$

RU_k represents the k unit in language rhythm, Mark (RU_k) represents the K rhythm mark and $Mark_j$ represents the class of language rhythm.

As before analysis, the marks in each kind of language rhythm are limited even very little. There are limited states need to be calculated in State Switch Matrix, so this method have great time and space performance.

4. Experiment

Choosing 600 documents and reports of three topics in three months and each topic contained about 200. Then mix 600 documents and reports together. Extract the language rhythm of each one to build the state switch matrix, classifying them with Bayes classifier. Each topic represents one class. Then the result of classifier can reflect the tracking performance.

4.1 Evaluating Indicator

There are two evaluating indicators: Recall and Precision.

Recall: The number ($RFlieS_i$) of the documents of one class obtained by classifier divided by the actual number(CFileSs) of the same class. as shown in formula 2 below:

$$\mathrm{Re}call_i = RFileS_i / CFileS_s \quad (i = 1, 2, \cdots n) \quad (2)$$

Precision: The proportion of the documents in correct class. As shown in formula 3 below:

$$\mathrm{Pr}ecision_i = \sum_{j=1}^{RFileS_i} Class(PFile_j) / RFileS_i \quad (3)$$

4.2 Experiment Process Analysis

The 600 documents about 3 topics in 3 months are classed by the classifier with the language rhythm feature, the result is shown as table1 and table 2.From the table1 we can see that most documents are in the right class. But 31 items that classifier detects to topic I are belong to topic III actually and the same with 12 to topic II and etc.

Table1 Result of classifier

	Topic I in classifier	Topic II in classifier	Topic III in classifier
Topic I in fact	163	21	16
Topic II in fact	9	172	19
Topic III in fact	31	12	157

And there are still a few documents cannot be classified correct. But most of them can be classified successfully. The recalls are 81.5%, 86% and 78.5% .The precisions are all above 80%.With the Table2 ,it is can be concluded that the language rhythm can get the good performance in Topic Tracking. The recalls and precisions are shown in table2.

Table2 Recall and Precision

Topic	Recall	Precision
Topic I	81.5%	80.3%
Topic II	86%	82.3%
Topic III	78.5%	81.78%
Average	82%	81.45%

5. Summary

The experiment has proved that language rhythm can complete the task of Topic Tracking with great performance. This method can find the documents and reports of the same topic among massive text. Because the sematic understanding is not necessary, the other more simple and effective feature is used. And it very appropriate to working in large scale documents. When people are talking the same topic, they are probably in the same thinking, emotion and reflected the same objective facts. So maybe they use different words to describe what they thinking, seeing and listening, but something hidden are sometimes the same, and this can be captured partly as language rhythm. As a result, language rhythm can be used as the feature of documents in Topic Tracking.

References

[1]. Yu Hong,Yu Zhang,Ting Liu,Sheng Li,Topic Detection and Tracking Review,Journal of Chinese Information Processing [J].2007, 11 (6):71-87.

[2]. S.A. Lowe, “The Beta-Binomial Mixture Model for Word Frequencies in Documents with Applications to Information Retrieval,” Proceedings of Euro speech ’99, Budapest, September 1999.

[3]. Luo Weihua, Liu Qun ,Development and Analysis of Technology of Topic Detection and Tracking Topic Detection and Tracking,Language Computing and Text Processing Based on Content--Institute of Computing Technology, Chinese Academy of Sciences [C].2003:560-566.

[4]. Francine Chen, AymanFarahat, Thorsten Brants, MultipleSimilarity Measures and Source-Pair Information in Story Link Detection [C]HLT-NAACL.2004:313-320

[5]. Peter D. Turney, Patrick Pantel From Frequency to Meaning:Vector Space Models of Semantics Journal of Articial Intelligence Research [J] 2010(37) :141-188

[6]. Victor Gonzalez-Castro, Rocio Alaiz-Rodriguez, Enrique Alegre, Class Distribution Estimation Based on the Hellinger Distance, Information Sciences[J], 2013 V(218):146–164

[7]. David M. Blei, Probabilistic Topic Models [J]. April 2012,vol. 55, no. 4 , Communications of the ACM 83

[8]. MartijnSpitters, Wessel Kraaij.Using Language Models for Tracking Events of Interest over Time[C]. Proceedings of LMIR 2001.Pittsburgh:[s.n.]2001:60-65