A Research on Automatic Chinese Catchword Extraction

Han Ren, Donghong Ji, Lei Han
Computer Science Department, Wuhan University, China
E-mail: cslotus@mail.whu.edu.cn, donghong_ji2000@yahoo.com.cn, hattason@126.com

Abstract

Catchwords refer to popular words or phrases within certain area in certain period of time. In this paper, we propose a novel approach for automatic Chinese catchwords extraction. At the beginning, we discuss the linguistic definition of catchwords and analyze the features of catchwords by manual evaluation. According to those features of catchwords, we define three aspects to describe Popular Degree of catchwords. To extract terms with maximum meaning, we adopt an effective ATE algorithm for multi-character words and long phrases. Then we use conic fitting in Time Series Analysis to build Popular Degree Curves of extracted terms. To calculate Popular Degree Values of catchwords, a formula is proposed which includes values of Popular Trend, Peak Value and Popular Keeping. Finally, a ranking list of catchword candidates is built according to Popular Degree Values. Experiments show that automatic Chinese catchword extraction is effective and objective in comparison with manual evaluation.

1. Introduction

Generally, a catchword, which is a term that represents a hot social phenomenon or an import incident, is paid attention by society public within certain period of time. For one thing, catchwords are a part of public focus, representing the mass value orientation. For another, they have a high timeliness. There are quiet a few evaluations every year. Only In 2005, tens of Chinese organizations published their ranking list of Chinese catchwords.

Being a language phenomenon, catchwords relate to important incidents and focused problems of current society as well as every aspect of our lives. Catchwords contain a great deal of information, a high timeliness and a wide spreading area, and they can truly and vividly reflect changes of our lives and our society. By monitoring, analyzing and researching catchwords, the changes of language rules can be found out, which can be as a reference to establish and adjust state language policy; and can be helped to language's healthy development and information security.

Currently, two kinds of approaches are adopted to evaluate catchwords. One is by CTR(Click-Through Rate) or retrieval times, but the limitation is that it is just based on frequency, which is only one part of catchwords' feature; the other is by experts evaluation, but it depends to a large extent on their subjective judgment.

In this paper, we propose a novel approach that can automatically analyze and extract Chinese catchwords. By analyzing evaluated catchwords and finding out their common feature called Popular Degree, we provide a method of popular degree quantification and give a formula to calculate term's popular degree value. After ranking, terms that have high values are picked out as candidate catchwords. The result can be provided as a reference for human evaluation. Experiment shows that automatic Chinese catchword extracting can promote the precision and objectivity as well as lighten the evaluation workload.

The rest of the paper is organized as follows. Section 2 discusses the theoretical basis of catchword judgment. Section 3 describes extraction in detail. In Section 4, the experimental results are shown as well as some discussions. Finally, a summary and conclusion is given in Section 5.

2. Theoretical basis of catchword Decision

The popularity of a word or phrase contains two factors: time and area, namely how long it lasts and how far it spreads. But neither of them have definite criterion. Thus the popularity should be defined clearly so that we can analyze catchword's feature.

2.1 Linguistic definition of catchword

Many researches of catchwords come from linguistics. Viewpoints exist in the academic circle but most of them are similar. Wang (1997) thought that catchwords, which include words, phrases, sentences or special patterns, are a language form in certain times and among certain groups or communities. Guo (1999) thought that catchwords are popular words, which are widely used in certain period of time among certain groups of people. To sum up definitions above, catchwords are a language form spreading quickly within certain area in certain period of time.

According to Zipf's Law (Zipf, 1949), the word that has a higher usage frequency is shorter than others. After analyzing catchwords evaluated, we find that they also follow this principle. Most catchwords are words and phrases instead of sentences and longer language units, which, at the same time, are difficult to extract automatically. In the paper, we focus catchwords on words and phrases.

2.2 Features of catchword

Some linguists do researches on features of catchwords but few of them propose quantification approaches to weigh features. Zhang (1999) proposes a method to judge catchwords by weighing Circulating Degree of catchwords, which are based on Dynamic Circulating Corpus. But the corpus construction and the judgment still

depend on manual evaluation.

By analyzing usage frequency of catchwords we find that, being a language phenomenon within a period of time, catchwords have two features; the one is high usage frequency, namely a catchword is largely used in certain period of time; the other is timeliness, namely this condition will lasts some time. Our quantification method is based on these features.

3. Automatic Chinese Catchword Extraction

In this section, we consider how to represent features of catchwords and how to evaluate these features by quantitative method.

3.1 Term Extraction

Catchwords that we need to extract are words or phrases with maximum meaning, most of which are multi-character words or long phrases. Word segmentation has a low discrimination for long phrases, while term extraction has a better way to extract them. Zhang (2006) proposes a new ATE algorithm, which is based on the decomposition of prime string. The algorithm evaluates the probability of a long string to be a term by weighing relation degree among sub-strings within the long string. The algorithm can raise the precision in extracting multi-character words and long phrases. In this paper, we use this method to extract catchwords.

3.2 Popular Degree Curve

For extracted terms, time granularity should be defined to describe their features. We select 'day' as the time granularity and get every day's usage frequency for each term in one year. These can be described as a time series like below:

$$C_w = \{c_{w1}, c_{w2}, \cdots, c_{wt}, \cdots, c_{wn}\} \quad (1)$$

C_w is the time series of term w, c_{wt} is the usage frequency of term w in the day t and n is the number of observational days. After getting usage frequency, we use SMA (Simple Moving Average) to deal with it that can eliminate random fluctuation. The formula is as follows:

$$\overline{c}_{wt} = \frac{\sum_{j=1}^{m} c_{w(t-m+j)}}{m} \quad (2)$$

$\overline{c}_{wt}$ is the smoothed usage frequency of term w in the day t and m is the interval. A short interval has little effect, while a long one may result in low accuracy. The proper interval is between 10 and 20. Smoothed time series is as follows:

$$\overline{C}_w = \{\overline{c}_{w1}, \overline{c}_{w2}, \cdots, \overline{c}_{wt}, \cdots, \overline{c}_{wn}\} \quad (3)$$

By analyzing smoothed time series curve we can see that, catchwords appear in certain period of time; in the period, its usage frequency change from low to high. After reaching the highest point, catchwords' usage frequency decrease slowly. We call this process Popular Degree, which can be described as three aspects:

1) **Popular Trend**: the changing process of usage frequency from low to highest; the more obvious the popular trend changes, the higher the popular degree is.

2) **Peak Value**: maximum usage frequency within certain period of time; the larger the peak value is, the higher the popular degree is.

3) **Popular Keeping**: the changing process of usage frequency from highest to low; the more gentle the popular keeping changes, the higher the popular degree is.

Three aspects above determine popular degree of catchwords. Figure 1 shows the smoothed time series curve of the catchword '苏丹红'[1] evaluated in year 2005:

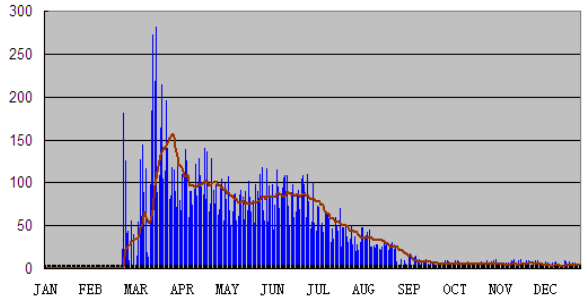

Figure 1: Smoothed time series curve of the catchword '苏丹红'

To the catchword '苏丹红', its Popular Trend changes obviously and its Popular Keeping changes gently. So the catchword '苏丹红' has a high Popular Degree.

Time series of catchwords belong to nonlinear trend. According to three aspects of Popular Degree, smoothed time series curve is separated into two parts: one is ascending period, namely Popular Trend process; the other is descending period, namely Popular Keeping process. We use conic fitting to deal with two parts of series data. The form of a conic's formula is as follows:

$$Y = a + bt + ct^2$$

According to least square method, a standard equations that can deduce three parameters a, b and c is as follows:

$$\begin{cases} \sum Y = na + b\sum t + c\sum t^2 \\ \sum tY = a\sum t + b\sum t^2 + c\sum t^3 \\ \sum t^2 Y = a\sum t^2 + b\sum t^3 + c\sum t^4 \end{cases}$$

Assume T_S is the starting time, T_E is the ending time, and T_M is the time that time series curve reaches the highest point. According to conic fitting method mentioned above we can get ascending and descending period curves. Formulas of two conics are as follows:

$$\begin{cases} \varphi(u) = a + bu + cu^2 & T_S \le t \le T_M \\ \psi(v) = a' + b'v + c'v^2 & T_M \le t \le T_E \end{cases} \quad (4)$$

Variable u and v are usage frequency of a day, $\varphi(u)$ is the formula of ascending curve, and $\psi(v)$ is the formula of descending curve. The curve described by equation (4) is called Popular Degree Curve. Figure 2 shows the Popular

[1] 苏丹红 means Sudan red in english.

Degree Curve of the catchword '苏丹红':

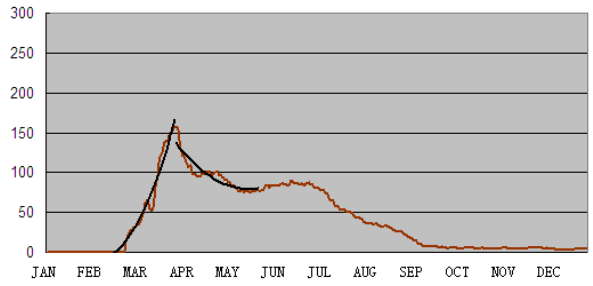

Figure 2: Popular Degree Curve of the catchword '苏丹红'

Figure 2 indicates that, being a trend curve, conic can be well used to describe Popular Trend and Popular Keeping process.

3.3 Popular Degree Value

The decision of catchwords is based on three aspects of Popular Degree. We propose a formula that can calculate Popular Degree values of catchwords. After getting the values, a ranking list from high scores to low ones is established. The Popular Degree of a catchword is in the direct ratio to its place in the ranking list. The formula is as follows:

$$PD(w) = PT(w) \times PV(w) \times PK(w) \tag{5}$$

$PD(w)$ is the Popular Degree value of the catchword w. $PT(w)$ is the Popular Trend value of w:

$$PT(w) = \alpha \bullet \frac{\varphi(T_M) - \varphi(T_S)}{\varphi(T_M)} \tag{6}$$

α is the adjusting parameter of Popular Trend. The formula indicates that $PT(w)$ is related to changing process of Popular Degree Curve from low to highest. $PV(w)$ is the Peak Value of w:

$$PV(w) = \beta \bullet \frac{\max\{\overline{c}_{wt}\}}{\frac{1}{N_w}\sum_w \max\{\overline{c}_{wt}\} + \max\{\overline{c}_{wt}\}} \tag{7}$$

β is the adjusting parameter of Peak Value. The formula indicates that $PV(w)$ is related to the maximum usage frequency of w. $PK(w)$ is the Popular Keeping value of w:

$$PK(w) = \gamma \bullet \left(1 - \frac{\psi(T_M) - \psi(T_E)}{\psi(T_M)}\right) \tag{8}$$

γ is the adjusting parameter of Popular Keeping. The formula indicates that $PK(w)$ is related to changing process of Popular Degree Curve from highest to low. Parameter α, β and γ control proportion of three aspects in Popular Degree Value.

All extracted terms are ranked according to their Popular Degree Values. Terms that have high scores are picked out as catchword candidates.

3.4 Algorithm

The algorithm of automatic catchwords extraction is described below:

Algorithm Extracting catchwords
Input text collections
Output ranking list of catchword candidates
Method
1) use ATE algorithm mentioned in section 3.1 to extract terms
2) filter terms that contains numbers and punctuations
3) **foreach** term
4) calculate its smoothed time series by formula (2)
5) use conic fitting method in section 3.2 to get its Popular Degree Curve like equations (4)
6) use formula (5) ~ (8) to calculate its Popular Degree value
7) rank all Popular Degree values from high to low

4. Experimental Results and Analysis

4.1 Text Collections

In the experiment, we use 136,191 web pages crawled from Sina's news reports [2] in 2005 including six categories: economy, science, current affairs, military, sports and entertainment. For the experiment purpose, we extract body content in every web page by using Noise Reducing algorithm (Shianhua Lin & Janming Ho, 2002). Totally, the extracted subset includes 129,328 documents.

4.2 parameter setting

Large time granularity may result in low accuracy for conic fitting. In this paper, we select 'day' as the time granularity and year 2005 as the range of statistical time. For the interval m in formula (2), a proper value should be specified not only eliminate random fluctuation but also keep accuracy of data. In the experiment we find the proper interval is between 10 and 20. We specify middle value of the range to m, namely m = 15.

Catchwords have a high timeliness, so we should specify an analyzing range of time. By analyzing catchwords evaluated, we find that Popular Trend process and Popular Keeping process both need approximately 3 months. So we specify the analyzing range of time is n / 2. Thus the relationship among the starting time T_S, the ending time T_E and the highest point time T_M are below:

$$T_S = T_M - \left\lceil \frac{n}{4} \right\rceil, \quad T_E = T_M + \left\lceil \frac{n}{4} \right\rceil$$

For keeping Popular Degree values of catchwords are in the range [0, 1], three adjusting parameters should be satisfied to the inequation: $0 < \alpha, \beta, \gamma \le 1$. In the experiment, we assume that three aspects of Popular Degree have the same effect on Popular Degree values of catchwords. So three parameters can be settled as 1. Table 1 shows all parameters involved in the experiment.

[2] Sina's website is http://www.sina.com.cn/

parameter	value
n	365
t	[1, 365]
m	15
T_s	$T_M - \lceil n/4 \rceil$
T_E	$T_M + \lceil n/4 \rceil$
α	1
β	1
γ	1

Table 1: parameters in the experiment

4.3 Experiment Results

We use algorithm described in section 3.4 to get the ranking list of catchword candidates. According to ATE algorithm mentioned in section 3.1, we extract 966,532 terms. After filtering invalid terms we get 892,184 terms and calculate each term's Popular Degree value. Table 2 shows top ten catchword candidates ranking list according to their popular degree values:

catchword candidates[3]	PD value
苏丹红	0.251262
超级女声	0.220975
油价	0.213843
纺织品谈判	0.196326
TD-SCDMA	0.185691
芙蓉姐姐	0.166730
发现号	0.154803
丁俊晖	0.137211
六方会谈	0.121738
猪链球菌	0.120667

Table 2: PD values of Top 10 catchword candidates

Currently, there is no unified standard for catchword evaluation. We compare the result with published catchwords evaluated by the Chinese Ministry of Education and NLRMRC (National Language Resources Monitoring and Research Centre). Figure 3 shows quantity of same catchwords extracted by our approach and artificial evaluation in both top N catchwords ranking list.

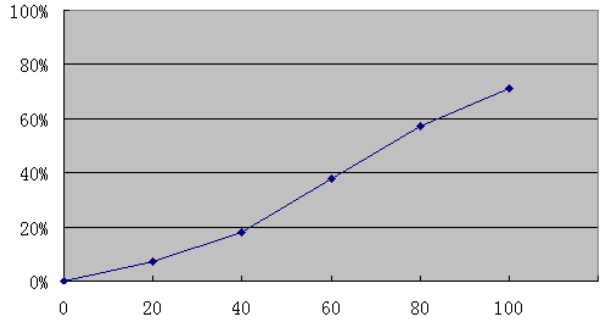

Figure 3: quantities of same catchwords between our approach and artificial evaluation in top N

Figure 3 indicates that, when N is 100, the ratio of same catchwords reaches over 70%; but when N is lower, the ratio is also lower. For one thing, we can see that our approach has a good effect on automatically extracting catchwords, closing to the result of evaluation by human. For another, it proves that subjectivity of catchwords evaluation greatly influences the result of ranking.

4.4 Analysis

In our experiment, Popular Values of some catchwords evaluated by organizations are lower. By analyzing their time series curves, we find that their usage frequencies are not high. We also find that these catchwords mostly have other expressions, such as the catchword '社会保障体系'[4], which can be also called '社保体系'[5]. These two synonyms are treated as one term in artificial evaluation that corresponds to promote usage frequency. However, relationship between the two synonyms is not concerned in our experiment. They are treated as separate terms. So the Popular Values of these two synonyms are not high either. It proves that parts of catchwords by artificial evaluation are collected and generalized; a catchword is considered not only a single word or a phrase, but also a part of a word-cluster which consist of synonymous words or phrases. Besides, dataset in our experiment is small, which influences the result of term extraction.

5. Conclusions

Being as one aspect of dynamic language research, catchword has a far-reaching significance for the development of linguistics. The paper proposes an approach that can automatically detect and extract catchwords. By analyzing evaluated catchwords and finding out their common feature called popular degree, the paper provides a method of popular degree quantification and gives a formula to calculate term's popular degree value. After ranking, terms that have high values are picked out as candidate catchwords. The result can be provided as a reference for catchword evaluation. Experiment shows that automatic catchword extraction can promote the precision and objectivity, and mostly

[3] 超级女声 means a talent show launched by Hunan Satellite.
油价 means petroleum price
纺织品谈判 means textile negotiation
TD-SCDMA means a third generation mobile communication standard
芙蓉姐姐 means a famous girl called sister lotus in Internet
发现号 means STS Discovery OV-103
丁俊晖 means a Billiards player named Junhui Ding
六方会谈 means Six-Party Talks
猪链球菌 means swine streptococcus suis

[4] 社会保障体系 means social security system
[5] 社保体系 is the abbreviation of 社会保障体系

lighten difficulties and workload of evaluation.
In the experiment, we also find that some catchwords are not isolated, but have a tightly relationship and express the same meaning. In the future, we can unite all synonymous catchwords to a word cluster and calculate the cluster's popular degree value. Thus we would be able to achieve a better performance for extraction.

6. Acknowledgements

This work is supported by the Natural Science Foundation of China (Grant No.60773011, 60703008).

7. References

G.E.P.Box, G.M.Jenkins and G.C.Reinsel. (1994). *Time Series Analysis, Forecasting and Control*. Third Edition, Prentice-Hall.

Richard L. Burden and J.Douglas Faires. (2001). *Numerical Analysis*. Seventh Edition, Brooks/Cole, Thomson Learning, Inc., pp. 186-226.

Xi Guo. (1999). *China Society Linguistics*. Nanjing : Nanjing University Press.

H. Kantz and T. Schreiber. (1997). *Nonlinear Time Series Analysis*. Cambridge University Press, 1997

Shianhua Lin, Janming Ho. (2002). *Discovering informative content blocks from Web documents.* In: SIGKDD.

Dechun Wang (1997). *Introduction to Linguistics*. Shanghai: Shanghai Foreign Language Education Press.

George K.Zipf (1949). *Human Behavior and Principle of Least Effort: an Introduction to Human Ecology*. Addison Wesley, Cambridge, Massachusetts.

Pu Zhang (1999). *On thinking of language sense and Circulating Degree*. Beijing: Language Teaching and Linguistic Studies, vol.(1).

Yong Zhang (2006). *Automatic Chinese Term Extraction Based on Decomposition of Prime String.* Beijing: Computer Engineering, vol.(23).