Automatic Chinese Catchword Extraction Based on Time Series Analysis

Han Ren[1], Donghong Ji[1], Jing Wan[2] and Lei Han[1]
1 School of Computer Science, Wuhan University 430079, China
2 Center for Study of Language & Information, Wuhan University 430072, China
cslotus@mail.whu.edu.cn, donghong_ji@yahoo.com,
jennifer.wanj@gmail.com, hattason@mail.whu.edu.cn

Abstract

Catchwords refer to those popular words or phrases in a time period. In this paper, we propose a novel approach for automatic extraction of Chinese catchwords. By analyzing features of catchwords, we define three aspects to describe Popular Degree of catchwords. Then we use curve fitting in Time Series Analysis to build Popular Degree Curves of the extracted terms. Finally we give a formula that can calculate Popular Degree values of catchwords and get a ranking list of catchword candidates. Experiments show that the method is effective.

1 Introduction

Generally, a catchword is a term which represents a hot social phenomenon or an important incident, and is paid attention by public society within certain time period. On the one hand, catchwords represent the mass value orientation for a period. On the other hand, they have a high timeliness. Currently, there are quiet a few ranking and evaluations of catchwords every year in various kinds of media. Only in year 2005, tens of Chinese organizations published their ranking list of Chinese catchwords.

Catchwords contain a great deal of information from any particular area, and such words truly and vividly reflect changes of our lives and our society. By monitoring and analysis of catchwords, we can learn the change of public attention in time. In addition, we may detect the potential changes of some linguistic rules, which can help establish and adjust state language policies.

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 feature of catchwords. The other is by manual evaluation, but it depends on their subjective judgment to a large extent. In this paper, we propose a novel approach that can automatically analyze and extract Chinese catchwords. By analyzing sample catchwords and finding out their common features, we provide a method to evaluate the popular degree. After ranking, terms that have high values are picked out as catchword candidates.

The rest of the paper is organized as follows. In Section 2, we discuss about the linguistic basis of catchword judgment. In Section 3, we describe the extraction method in detail. In Section 4, we present the experimental results as well as some discussions. Finally, we give the conclusion and future work in Section 5.

2 Linguistic basis

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.

2.1 Linguistic definition of catchword

Many researches of catchwords come from pure linguistic areas. Wang (1997) proposed 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) specified that catchwords are popular words, which are widely

CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pages 113–118
Manchester, August 2008

used in certain period of time among certain groups of people. To sum up, 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. Catchwords also follow this principle: most catchwords are words and phrases instead of sentences and longer language units, which are more difficult to extract automatically. In the paper, we focus on catchwords as words and phrases.

2.2 Features of catchword

Some features of catchwords have been proposed, but there have been few research to quantify and weigh the features. Zhang (1999) proposed 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 efforts.

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

3 Extraction Method

In this section, the extraction method is described in detail. After term extraction, the features of terms are weighed by time series analysis. The algorithm in section 3.4 shows the process to extract catchword candidates.

3.1 Term Extraction

Catchwords are words or phrases with maximal meanings, most of which are multi-character words or phrases. Word segmentation has a low discrimination for long phrases, while term extraction has a better way to extract them. Zhang (2006) proposed 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 terms.

3.2 Popular Degree Curve

For extracted terms, a 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:

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

C_w is the time series of term *w*. c_{wt} is the usage frequency of term *w* in the day *t*. *n* is the number of observation days.

As a latent knowledge, two features of catchwords mentioned in section 2.2 exist in their time series. The effective method to find out the latent knowledge in the time series is Time Series Analysis, which includes linear analysis and nonlinear analysis. As the time series of terms belong to nonlinear series, we use nonlinear analysis to deal with them.

After getting usage frequency, we use SMA (Simple Moving Average) method to eliminate the random fluctuation of series C_w . 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. In SMA method, a short interval has a little effect, while a long one may result in low accuracy. So we should specify a proper interval. Through experiments we find that an appropriate interval is between 10 and 20. Smoothed time series is as follows:

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

Smoothed time series of terms can be described as curves, in which the coordinate *x* is day *t* and coordinate *y* is $\overline{c}_{wt}$. Through these curves we can see that, catchwords appear in certain period of time and its usage frequency increases in this period. After reaching the highest point, usage frequency of catchwords decrease slowly. We call this process Popular Degree, which contains three aspects:

1) Popular Trend: the increasing process of usage frequency; the more obviously 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 decreasing process of usage frequency; the more gently 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[2]', evaluated in year 2005:

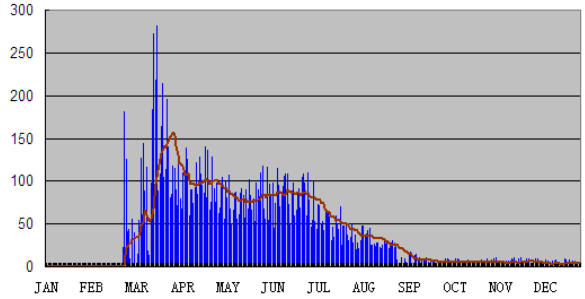

Figure 1. Smoothed time series curve of the catchword '苏丹红'

To the catchword '苏丹红', its Popular Trend changes obviously and its Popular Keeping changes gently. Meanwhile, its Peak Value is relatively higher than those of most catchwords. So the catchword '苏丹红' has a high Popular Degree.

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. A conic's formula is like below:

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

According to least square method, a standard equation 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 we can get curves of ascending and descending period. 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 term in 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 Degree Curve of the catchword '苏丹红':

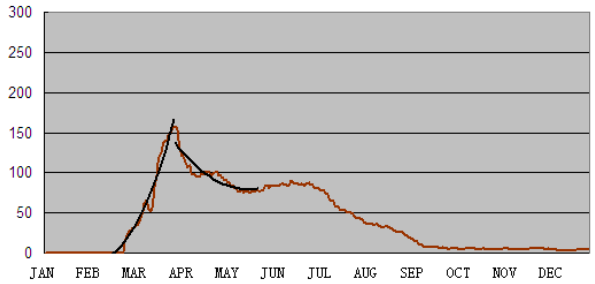

Figure 2. Popular Degree Curve of the catchword '苏丹红'

3.3 Popular Degree Value

The decision of catchwords is based on three aspects of Popular Degree described in section 3.2. We propose a formula to calculate Popular Degree values of terms. After getting the values, a ranking list by inverse order 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:

$$PD(w) = PT(w) \times PV(w) \times PK(w) \quad (5)$$

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

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

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

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

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:

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

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

[2] 苏丹红 means Sudan red in English.

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 equation (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 Collection

In the experiment, we use 136,191 web pages crawled from Sina[3]'s news reports in year 2005 including six categories: economy, science, current affairs, military, sports and entertainment. For the experimental 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 Experiment settings

In the experiment, several parameters should be settled to perform the catchwords extraction.

• n

A large time granularity may result in low accuracy for conic fitting. In this paper, we select 'day' as the time granularity.

• m

For the interval m in formula (2), a proper value should be specified to not only eliminate random fluctuation but also keep accuracy of data. In the experiment we find that the proper interval is between 10 and 20.

• T_S and T_E

Catchwords have a high timeliness, so we should specify a time domain. By analysis of sample catchwords, we find that popular time domain for most of them approximately last for not more than 6 months. So we specify the time domain is n / 2. Thus the relationship among the starting time T_S and the ending time T_E is below:

$$T_S = T_E - \frac{n}{2}$$

As a proper example, the starting point can be 60 days away from the highest point. Thus the Popular Trend process and the Popular Keeping process both last for nearly 3 months. So the relationship can be described as formulas below:

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

• α, β, γ

To keep the Popular Degree values of catchwords within [0, 1], three adjusting parameters are satisfied to the inequation: $0 < \alpha, \beta, \gamma \le 1$.

Table 1 shows proper values of parameters as schema 1. We also give other schemas, which contain different values of parameters, to compare with the schema 1. In schema 2 to schema 4, default values of parameters are the same with schema 1.

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 schema 1

schema 2: different m values
schema 3: different values of T_S and T_E
schema 4: different values of α, β and γ

4.3 Evaluation Measure

Currently, there is no unified standard for catchword evaluation. In year 2005, NLRMRC

[3] http://www.sina.com.cn/

(National Language Resources Monitoring and Research Centre, held by MOE of China) had published their top 100 Chinese catchwords. We use co-occurrence ratio of catchwords for the evaluation. The formula of co-occurrence ratio is as follows:

$$r = \frac{N_C}{N}$$

N is the number of ranking catchwords. N_C is the co-occurrence of catchwords, namely the number of catchwords which appear both in our approach and NLRMRC in top N.

4.4 Results

We use algorithm described in section 3.4 to get a 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 - 5 shows the co-occurrence ratio with schema 1 - 4.

N=20	N=40	N=60	N=80	N=100
7%	18%	36%	53%	66%

Table 2. Co-occurrence ratio using schema 1

m	N=20	N=40	N=60	N=80	N=100
5	3%	7%	16%	29%	45%
10	4%	11%	25%	44%	59%
20	**7%**	**15%**	**32%**	**49%**	**63%**
25	6%	14%	29%	46%	60%

Table 3. Co-occurrence ratio using schema 2

T_M - T_S : T_E - T_M	N=20	N=40	N=60	N=80	N=100
1 : 4	0%	3%	8%	15%	22%
2 : 3	4%	14%	30%	49%	**64%**
3 : 2	**5%**	**15%**	**33%**	**51%**	63%
4 : 1	2%	5%	12%	21%	26%

Table 4. Co-occurrence ratio using schema 3

	N=20	N=40	N=60	N=80	N=100
α=0.5	3%	9%	24%	42%	55%
α=0.8	**6%**	**15%**	31%	50%	**64%**
β=0.5	2%	6%	16%	37%	52%
β=0.8	5%	13%	29%	47%	59%
γ=0.5	3%	11%	26%	43%	57%
γ=0.8	**6%**	**15%**	**32%**	**51%**	62%

Table 5. Co-occurrence ratio using schema 4

Table 2 shows the co-occurrence ratio of the catchwords extracted by our approach and NLRMRC in top N catchwords ranking list. It indicates that, when N is 100, co-occurrence of the catchwords reaches 66%; when N is lower, the ratio is also lower. On the one hand, we can see that our approach has a good effect on automatically extracting catchwords, closing to the result of manual evaluation with the increment of N. On the other hand, it proves that divergence exists between our approach and manual evaluation in high-ranking catchwords.

Table 3 indicates that, the condition of m = 20 has a better co-occurrence ratio in contrast with others in schema 2. It is because a short interval has a little effect, while a long one may result in low accuracy in SMA.

Table 4 indicates that a better performance can be made when the proportion of T_M - T_S and T_E - T_M is close to 1:1. It proves that Popular Trend process is just as important as Popular Keeping process. Therefore the best time domain of these two processes are both n / 4.

Three parameters can adjust the weights of *PD*, *PV* and *PK* in formula (5). Table 5 indicates that three factors above are all important for weighing a catchword, while β is a little more important than α and γ. Therefore, maximum usage frequency of a catchword is a little more important than two other factors.

From Table 2 – 5 we can see that, parameters in schema 1 is most appropriate for the evaluation.

Table 6 shows the ranking list of top 10 catchword candidates according to their Popular Degree values:

candidates[4]	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 6. Popular Degree values of Top 10 catchword candidates

[4] 超级女声 means a talent show by Hunan Satellite.
油价 means petroleum price
纺织品谈判 means textile negotiation
芙蓉姐姐 means a famous girl called sister lotus
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

4.5 Analysis

In our experiment, Popular Values of some catchwords by manual evaluation are lower. By analyzing their time series curves, we find that usage frequencies of these terms are not high. We also find that these catchwords mostly have other expressions. Such as the catchword '社会保障体系[5]' can be also called '社保体系[6]'. These two synonyms are treated as one term in manual evaluation that corresponds to promote usage frequency. However, relationship between the two synonyms is not concerned in automatic extraction. They are treated as separate terms. So the Popular Degree Values of these two synonyms are not high either. It proves that parts of catchwords by manual evaluation are collected and generalized. A catchword should be treated not only as a separate word or a phrase, but also as a part of a word-cluster, which consist of synonymous words or phrases. Through word clustering method, we can get an increasing quantity of the co-occurrence of catchwords between our approach and manual evaluations.

5 Conclusions

Being as one aspect of dynamic language research, catchwords have 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 catchword candidates. The result can be provided as a reference for catchword evaluation. Experiments show that automatic catchword extraction can promote the precision and objectivity, and mostly lighten difficulties and workload of evaluation.

In the experiment, we also find that some catchwords are not isolated, but have a strong 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.

Acknowledgement

This work is supported by the Natural Science Foundation of China under Grant Nos.60773011, 60703008.

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, (1).

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

[5] 社会保障体系 means social security system

[6] 社保体系 is the abbreviation of 社会保障体系