A Knowledge-based Approach to Text Classification

Zhu Jingbo
Institute of Computer Software & Theory
Northeastern University, Shenyang Liaoning,
P.R.China 110006
zhujingbo@yahoo.com

Yao Tianshun
Institute of Computer Software & Theory
Northeastern University, Shenyang Liaoning,
P.R.China 110006
tsyao@china.com

Abstract

The paper presents a simple and effective knowledge-based approach for the task of text classification. The approach uses topic identification algorithm named FIFA to text classification. In this paper the basic process of text classification task and FIFA algorithm are described in detail. At last some results of experiment and evaluations are discussed.
Keywords: FIFA algorithm, topic identification, text classification, natural language processing

Introduction

The text automatic classification method is based on the content analysis automatically to allocate the text into pre-determined catalogue. The methods of text automatic classification mainly use information retrieval techniques. Traditional information retrieval mainly retrieves relevant documents by using keyword-based or statistic-based techniques (Salton.G1989). Generally, three famous models are used: vector space model, Boolean model and probability model, based on the three models, some researchers brought forward extended models such as John M.Picrrc(2001), Thomas Bayer, Ingrid Renz,Michael Stein(1996), Antal van den Bosch, Walter Daelemans, Ton Weijters(1996), Manuel de Buenaga Rodriguez, Jose Maria Gomez-llidalgo, Belen Diaz-agudo(1997), Ellen Riloff and Wendy Lehnert(1994).

One central step in automatic text classification is to identify the major topics of the texts. We present a simple and effective knowledge-based approach to text automatic classification. The approach uses topic identification algorithm named FIFA to text classification. In this paper the basic process of text classification task and FIFA algorithm are described in detail. At last some results of experiment and evaluations are discussed.

1 Knowledge-based text classification

The principal process for the Knowledge -based text classification is illustrated as following:

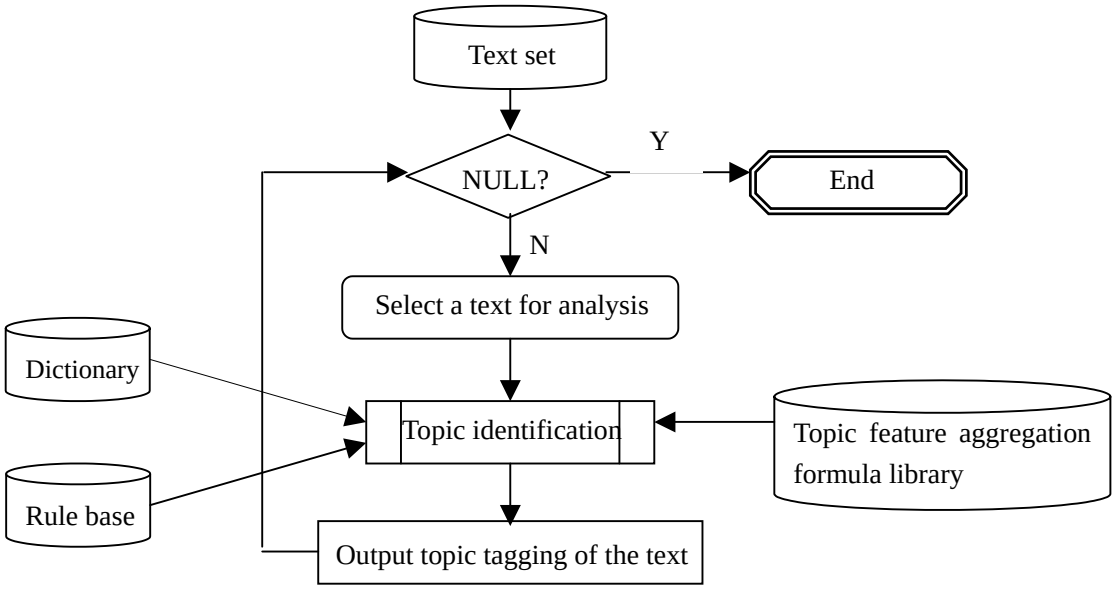


Figure 1 The principal process for the knowledge-based text classification

From the figure 1 we can know that the crucial technique of the text classification is the topic identification parser. The topic tagging of the text is identified as its catalogue.

2 Automatic Topic Identification

2.1 Feature Dictionary

The feature dictionary is mainly used to store some terms that can illustrate the topic feature concept, and we call these terms as "feature terms". The data structure of the feature dictionary is consist of word, POS, semantic, location and field attribute. Some examples of feature dictionary are described as following:

Feature terms	Attributes
会同县 HuiTong County	S(),,地名,地区(会同县),,国名(中国)省市(湖南)地区(怀化),领域(地理) S(),,LOCATION,LOCATION(HuitongCounty),,COUNTRY(China)PROVINCE (HuNan)CITY(HuaiHua),FIELD(geography)
中国银行 Bank of China	N(),,银行,机构(工商银行),,国名(中国),领域(银行)领域(金融) N(),,BANK,ORGANIZATION(bank),,COUNTRY(China),FIELD(bank) FIELD(finance)
作战情报中心 battle information center	N(),,情报中心,机构(作战情报中心),,,领域(情报)领域(军事) N(),,INFORMATION-CENTER,ORGANIZATION(battle information ceter),,,FIELD(information)FIELD(military)
三峡工程 Sanxia Project	N(),,工程,工程(三峡工程),,国名(中国),领域(水利) N(),,PROJECT,PROJECT(Sanxia Project),,COUNTRY(China), FIELD (irrigation)

Table 1: Some examples in the feature dictionary

Since 1996 we employed a semi-automatic method to acquire feature terms from pre-categorized corpus, and developed a feature dictionary including about 300,000 feature terms. There are about 1,500 kinds of semantic features, and about 1,000 kinds of field attributes to tag feature terms in this dictionary.

2.2 Topic Feature Distribution Computing Formula

According to the field attributes, frequencies and positions of feature terms, we could compute topic feature distribution. The computing steps are described as following:

1) According to the frequency and position of a feature term ft_i, we could compute the weight of the term ft_i. The computing formula is described as following:

$$p(ft_i) = \frac{freq(ft_i) + N_{title} + 0.5 \times N_{begin} + 0.5 \times N_{end}}{\sum freq(ft_i)} \quad (1)$$

Where $p(ft_i)$ is the weight of the feature term ft_i. $freq(ft_i)$ is frequency of the feature term ft_i. N_{title} is times of the feature term ft_i occurring in the title. N_{begin} is times of the feature term ft_i occurring in the first sentence of a paragraph. N_{end} is times of the feature term ft_i occurring in the end sentence of a paragraph. $\sum freq(ft_i)$ is total frequency of the all feature terms in the text.

In the experiment we discovered that the feature terms in different position of a text have the different influence abilities on the topic features. So we take into account of this factor and use different experience coefficient in the weight computing formula of feature terms. In the formula (1), the coefficient of N_{title} is 1.0, the coefficient of N_{begin} is 0.5, and the coefficient of N_{end} is 0.5.

2) From the attribute of a feature term in the dictionary we could acquire its field attribute, in fact this field attribute is the topic feature illustrated by the feature term. The weight of a topic feature could be gotten by adding all weights of feature terms that illustrate the same topic feature. The more the feature terms illustrates the same topic feature, the higher the weight of the topic feature. The weight of a topic feature expresses its abilities illustrating the topic of the text. The weight $p(f_i)$ computing formula of a topic feature f_i is described as following:

$$p(f_i) = \sum_{ft_j \in f_i} p(ft_j) \quad (2)$$

Where $p(f_i)$ is the weight of the topic feature f_i. $p(ft_j)$ is the weight of the feature term ft_j. $ft_j \in f_i$ shows feature term set illustrating the same topic feature f_i.

2.3 Topic Feature Aggregation Formula

The topic feature aggregation formula is described as following:

$$\xi_i : \beta(t_i) = \sum_{j=1}^{n} p(f_j) \times \mu(f_j) \qquad (3)$$

Where ß(t_i) is the weight of the topic t_i. f_j is the topic feature illustrating the topic t_i. $p(f_j)$ is the weight of the topic feature f_j. μ (f_j) is the coefficient of the topic feature f_j.

In the application system, we used automatic construction technique to construct a library, which called topic feature aggregation formula library that includes 105 topic feature aggregation formulas.

2.4 FIFA algorithm

Most of automatic text processing techniques uses topic identification as part of a specific task. The approaches to topic identification taken in these techniques can be summarized in three group: statistical, knowledge-based, and hybrid. The statistical approach (H.P.Luhn 1957， H.P.Edmundson 1969， Gerard Salton, James Allan, Chris Buckley, and Amit Singhal 1994) infers topics of texts from term frequency, term location, term co-occurrence, etc, without using external knowledge bases such as machine readable dictionaries. The knowledge-based approach (Wendy Lehnert and C. Loiselle 1989, Lin Hongfei 2000) relies on a syntactic or semantic parser, knowledge bases such as scripts or machine readable dictionaries, etc., without using any corpus statistics. The hybrid approach (Elizabeth D. Liddy and Sung H. Myaeng 1992, Marti A. Hearst 1994) combines the statistical and knowledge-based approaches to take advantage of the strengths of both approaches and thereby to improve the overall system performance.

This paper presents a simple and effective approach named FIFA (feature identification and feature aggregation) to text automatic topic identification. The core of algorithm FIFA is based on the equation:

Topic Identification = Topic Feature Identification + Topic Feature Aggregation.

Topic identification (TI) can be divided into two phases: topic feature identification (FI) and topic feature aggregation (FA).

1) Topic Feature Identification (FI): We use the term 'topic feature' to name the sub-topic in a text. In this phase algorithm FIFA identifies feature terms[1] in a text by dictionary-based and rule-based methods. The distribution of a topic feature is computed by attributes, frequencies and positions of topic feature terms.

2) Topic Feature Aggregation (FA): According to distribution of topic features, in this phase we use topic feature aggregation formulas to compute the weights of topics of a text, then the topic of a text could be determined by the weights computed. Topic feature aggregation formula will be introduced detailedly in the following chapters. Using machine-learning method, the topic feature aggregation formulas could be acquired automatically from pre-classified training corpus.

The topic identification algorithm FIFA could be described as following:

Step1: Text segmentation and POS tagging

Input: a raw text

1. Preprocessing phase: One major function is to recognize sentence boundaries, paragraph breaks, abbreviations, numbers, and other special tokens.
2. Segmentation phase: Employing maximal matching algorithm to segment a sentences into some words, and setting a word's POS set in machine readable dictionary as its POS tagging.
3. Disambiguation phase: Employing a technology based on ambiguous segmentation dictionary[2] to resolve the problem of word ambiguous segmentation, and base on rules to recognize the unknown words, such as name, location, company, organization noun etc.
4. POS tagging phase: Employing tri-gram based technology to POS tagging.

Output: a text with formats, segmentation and POS tagging

Step2: Topic feature identification

Input: a text with formats, segmentation and POS tagging

1. Feature-dictionary-based feature terms identification and tagging

   The core of the method is to use feature dictionary to realize the feature terms identification and tagging. If a term in the text is found in the dictionary, then we call this term as a feature term of the text and its

[1] Perhaps is a word, phrase, string, etc. According to need of the application system to determine the type of the feature term.

[2] Is a machine readable man-made dictionary which includes examples of ambiguous segmentation and its correct segmentation

field attribute in the dictionary is tagged as the topic attribute of the feature term.

2. Rule-based feature terms identification and tagging

   Because of the limitation of the feature dictionary, we could not identify all feature terms by feature-dictionary-based technique. To resolve the problem of the unknown feature terms, we use the technique of rule-based feature terms identification and tagging. There are two steps for the identification and tagging:

   1) We employ statistics-based approach to acquire some high-frequency terms from the text as analysis objects which length is composed of two or more words, and the frequency in the text should exceed two times.

   2) We employ rule-based technique to analyze the grammatical structure of the high frequency terms, and according to the grammatical structure of the terms and the attribute of the central word to estimate the field attribute of the term, which is tagged as the topic feature attribute of the term.

3. According to attributes, frequencies and positions of the feature terms to calculate the distribution of the topic feature of the text.

   Output: topic feature set $\Psi(=\{(f_i,ß_i)\})$ of the text. Where f_i is the i^{th} text topic feature; $ß_i$ is the weight of the i^{th} text topic feature f_i subjected to $ß_i \in (0,1)$.

Step3: Topic feature aggregation

Input: The topic feature set Ψ of the text.

1. Reading a formula ξ_i from the topic feature aggregation formula library, where the formula ξ_i is the aggregation formula of the topic t_i.
2. According to parameters in the topic feature set Ψ, the weight of the topic t_i could be computed by the formula ξ_i.
3. If there are some other aggregation formulas in the library, then go to Step1, otherwise go to the next step.
4. Supposing the topic feature f_i in the set $\Psi(=\{(f_i,ß_i)\})$ as a topic, and $ß_i$ is the weight of the topic f_i. We sort the topic t_i and f_i by weight.

Output: Selecting the topic with maximal weight as topic tagging of the text.

Algorithm1: Topic identification algorithm FIFA

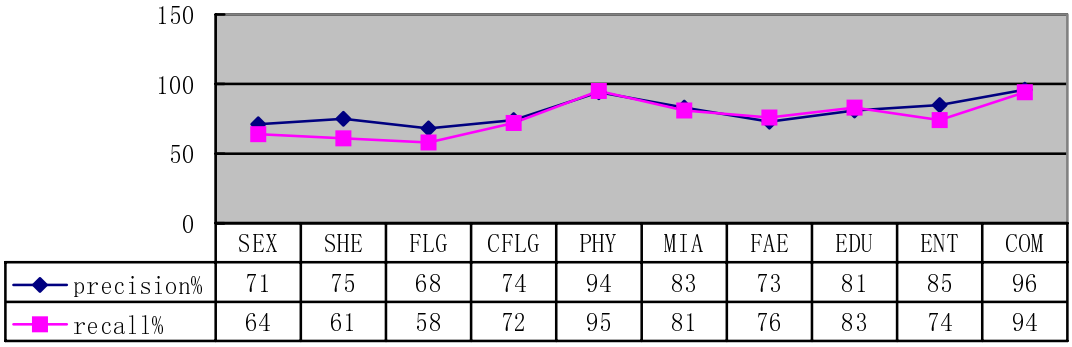
3. Experiment

To test the efficiency of the FIFA-based text automatic classification, and according to the pre-determined 10 topics we constructed a test corpus, which includes 1000 articles downloaded from the Internet. The composing of the test corpus is described as following:

Topic (abbreviation)	Number of articles
Sex (SEX)	100
Sex Healthy (SHE)	100
Fa Lun Gong (FLG)	100
Critical of Fa Lun Gong (CFLG)	100
Physical (PHY)	100
Military affairs (MIA)	100
Finance and economics (FAE)	100
Education (EDU)	100
Entertainment (ENT)	100
Computer (COM)	100
Total	1000

Table 2 The composing of the test corpus

Experiment 1: By classifying the test corpus, we could value the effect of the FIFA-based text automatic classification. The following figure 2 shows the results of text classification. Line ◆ represents precision percent while line ■ represents recall percent.



	SEX	SHE	FLG	CFLG	PHY	MIA	FAE	EDU	ENT	COM
precision%	71	75	68	74	94	83	73	81	85	96
recall%	64	61	58	72	95	81	76	83	74	94

Figure 2 The results of FIFA-based text classification

4. Conclusion

This paper presented a simple and effective approach to topic automatic identification. We use the topic identification approach for the task of text classification. The results of experiment show that a good precise and recall percent are achieved. In fact the topic identification approach called FIFA could be used not only as a stand-alone topic identification unit, but also in other text processing tasks such as text summarization, information retrieval, information routing etc.

References

Salton.G(1989), *Automatic Text Processing : The Transformation : Analysis and Retrieval of Information by Computer*, Addison-Wesley, Reading, Mass

John M.Picrrc(2001), *On the automated classification of web sites*, Linkoping Electronic Articles in Computer and Information Science, Vol.6, 2001, Sweden,

Thomas Bayer, Ingrid Renz,Michael Stein, Ulrich Kressel(1996), *Domain and language independent feature extraction for statistical text categorization*, proceedings of workshop on language engineering for document analysis and recognition - ed. by L. Evett and T. Rose, part of the AISB 1996 Workshop Series, April 96, Sussex University, England, 21-32 (ISBN 0 905 488628)

Antal van den Bosch, Walter Daelemans, Ton Weijters(1996), *Morphological analysis as classification: an inductive-learning approach*, Proceedings of NEMLAP-2, 2, July, 1996

Manuel de Buenaga Rodriguez, Jose Maria Gomez-llidalgo, Belen Diaz-agudo(1997), *Using WORDNET to complement training information in text categorization*, Second International Conference on Recent Advances in Natural Language Processing, 1997

Ellen Riloff and Wendy Lehnert(1994), *Information Extraction as Basis for High-precision Text Classification*, ACM Transactions on Information System, Vol12, No.3, July 1994

H.P.Luhn(1957). *A statistical approach to mechanized encoding and searching of literary information*. IBM Journal, p309-17, October 1957

H.P.Edmundson(1969). *New methods in automatic extracting*. Journal of the ACM, 16(2):264-85,1969

Gerard Salton, James Allan, Chris Buckley, and Amit Singhal(1994). *Automatic analysis, theme generation, and summarization of machine-readable texts*. Science, 264:1421-26, June 1994

Wendy Lehnert and C. Loiselle(1989). *An introduction to plot unit*. In David Waltz, editor, Semantic Structures-Advances in Natural Language Processing, p88-111, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989

Lin Hongfei(2000), *Logic Model for Chinese Text Filtering*, Ph.D dissertation, Northeastern University, 2000.3

Elizabeth D. Liddy and Sung H. Myaeng(1992). *DR-LINK's linguistic- conceptual approach to document detection*. In Proceedings of the First Text Retrieval Conferece (TREC-1), p113-29, 1992

Marti A. Hearst(1994). *Context and Structure in Automated Full-Text Information Access*. PhD thesis, Computer Science Division, University of California at Berkeley, California, April 1994