Memory Principle Based Domain Word Relationship Network Construction Model

Lei Wang* **Kuanjiu Zhou*** **Peng Qiu***

*Institute of Systems Engineering,
Dalian University of Technology, 116024, China
{wl.dut,zkj.dut,qp.dut}@163.com

Abstract

Automatic domain knowledge extraction is one of the key problems in text processing. To construct an effective domain word relationship network is important to text mining, text clustering and text information retrieving. A segmentation algorithm based on famous Apriori algorithm is issued and then a new construction model of domain word relationship network is proposed to organize domain words occurred in the same context with human memory principle. Some experiment results show that this model can extract domain words and construct their relationships from a great deal of domain documents effectively.

Keywords : memory model, automatic domain words extraction, word relationship network, Apriori algorithm

1. Introduction

As an important resource of knowledge, unstructured texts play a crucial role in human communication. Recently, information overload caused by the ever-increasing number and complexity of issues and insufficient support for information organization is becoming more and more serious. With the development of Chinese information processing and the increase of knowledge demand, text classification and filtering becomes an urgent issue.

In order to build a foundation for semantic analysis of text, much effort has been contributed to this area by researchers. Some of the most popular methods are implemented and evaluated using WordNet as the underlying reference ontology[1]. By Contrast, FrameNet introduces frame semantic to facilitate text understanding, which describes each frame's basic conceptual structure, gives names and descriptions for the elements which participate in such structures[2]. SOM is a powerful tool to settle the problem of semantic word clustering[3], but it requires word segmentation be finished before it takes place. In a word, the above methodologies are all based on human knowledge. In this paper, we present a prior-knowledge free method to mine word relationship, which gives a new way to explore the organization of words and build the foundation for further analysis of word sense. In addition, relevant researches on memory principle are also the foundation of this paper.

German philosopher Hermann Ebbinghaus completed his systematic researches on forgetting phenomenon in 1885, and Ebbinghaus Curve was painted from experimental data which used non-meaning syllables as memory material. He pointed out that the forgetting progress subjects to the constraint of other factors besides time. The firstly forgotten materials are insignificant, uninteresting things, which are unwanted to

people. Unfamiliar materials are easier to be forgotten than familiar ones. Since then various researches on memory science boomed rapidly, including researches on memorization of English vocabulary which give a great illumination to this paper.

According to the above analysis, a memory model is proposed to imitate learning process of brain of human being and a word relationship network from a certain domain texts is established. Firstly, domain words from large-scale text training documents are extracted and then relationships between words occurring in the same context are established according to the memory model. Experiments results show that this model can filter some irrelevant words effectively and the gained word relationship network reveales satisfactory quality.

2 Main flow of domain word relationship network construction

A construction process of the domain word relationship network is briefly depicted by the Fig. 1 as below:

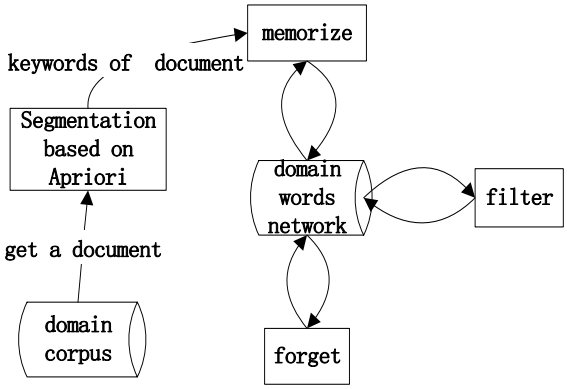

Fig.1. high level process for domain word relationship network construction

(1) Economic part of Fudan University corpus is used as domain training documents, which includes 1600 documents and provides plentiful domain words.

(2) Apriori-based segmentation algorithm is used to segment terms in each document. It is a statistic-based method which can extract terms with high frequencies as keywords of each document without dictionary support.

(3) Establishment of word relationship network.

Keywords of the same document are linked together according to memory model. Memorization and forgetting process are also introduced respectively in this model. Through imitating the learning process of human being brain, domain word relationship network will be constructed.

3 Apriori-based segmentation algorithm

Before introduction of the new algorithm, some definitions should be presented as follows:

Definition 1: *ItemSet*: a set used to store keywords and their corresponding occurrence frequency, *k-ItemSet* represents that the length of each keyword in *k -ItemSet* is *k*.

Definition 2: *Filtering*: a process that removes the terms whose occurrence frequency are less than a given threshold δ in *k-ItemSet*. The value of δ is set at 1 in this paper.

Definition 3: *Eliminate (Str, t)*: suppose a string Str's length as h, *Eliminate (Str, t) indicates that* subtract t from the occurrence frequency of the string Str in h-ItemSet.

The process of Apriori-based segmentation algorithm is presented here:

Input: *Text*: a document
StopList: including punctuation, auxiliary word etc
β: keywords proportion of all words extracted from the document

Output: *Keywords of the document*

Apriori-based segmentation algorithm:

Initialization

β = 0.05; sum = 0; //initial number of all distinct

words extracted from text document.

key = 0; // initial number of keywords of the document

Step 1: Count the occurrence frequencies of each character in *Text*, save the character and its occurrence frequency in *1-ItemSet* if it is not existed in StopList and then *Filter 1-ItemSet*;

Step 2: Each two adjacent characters in *Text* are extracted to form a word. If the word is not in *StopList* and both of the two characters in this word exist in the *1-ItemSet*, it would be added into *2-ItemSet* (only occurrence frequency value added if this word has already existed in *2-ItemSet*, else add it as a new word into *2-ItemSet* and initialize its occurrence frequency as 1). Then *Filter 2-ItemSet*;

Step 3: Create *k+1-ItemSet* from *k-ItemSet*

Choose two words in *k-ItemSet.*

$C=c_1c_2c_3...c_k$, $D=d_1d_2d_3...d_k$

$$If \ \forall i \in [2,k], \ c_i = d_{i-1}$$

then they can be connected to from a new word $(c_1c_2c_3...c_k \ d_k)$ as a candidate. After all candidates are generated, count their occurrence frequency in Text, save each word and its occurrence frequency in *k+1-ItemSet*, then *Filter k+1-ItemSet*;

Step 4: if *k+1-ItemSet* is not null, turn to step 3; else turn to step 5;

Step 5: Frequency Reduction

After the word has been extracted, all its sub-patterns, whether valid or invalid, may also be extracted, which could potentially increase the errors in word extraction. Therefore frequency reduction should be carried out to ensure that the invalid sub-patterns will not survive.

Eliminate from k-ItemSet and compute every sub-patterns of each words in k-ItemSet, and then use these sub-patterns to eliminate its occurrence frequencies in its corresponding ItemSet. After that k=k-1, continue eliminating as the above process until k=2 and finally all Itemsets should be filtered.

Step 6: keywords generation

After segmentation, all words in Itemsets are available. Sort these words according to its frequency. Count all distinct words: sum, and then several frequent words will be selected as keywords of the document. The amount of keywords is dependent on the following function: key=(int)sum*β .

4 Memory Model

Modern psychologists pointed out that input information can only be transformed into short-time memory on condition that they are paid attention to. However, only after encoding, short memory can be transformed into long-time memory and then be memorized, otherwise it will be forgotten. In a word, whether input information can be memorized is dependent on how your brain encodes this information [4]. Wang Jinmei indicates that according to memory principle, vocabulary memorization is influenced by the way it is encoded. Encoding is also a procedure of applying new vocabularies learning strategy. So teachers adopt a kind of clustering method to strengthen learning, which bind domain words together to accelerate learning process. In this way, information is interconnected, which will facilitate transformation from short-time memory to long-time memory and alleviate forgetting [4]. Domain word relationship network is organized just using this memory principle.

In this paper, we take every document as a scene in which a few keywords will appear concurrently, and then connection between each couple of two words will be established. When two words appear concurrently in other scene, their connection will be strengthened, this process called “memorization”. However when a new scene occurred, two words failed to

appear concurrently in this scene, their connection coefficient will decline, this process called “forgetting”.

According to memory principle, this paper trains the memory with sigmoid function, which function and inverse function are y=f(x) and x=f(y) respectively. These two functions are shown as below;

$$y = f(x) = \frac{1}{1+e^{-x}} \qquad x \in (-\infty, \infty)$$

$$x = f(y) = \ln(\frac{y}{1-y}) \qquad y \in (-1,1)$$

In the two functions, y denotes link coefficient between two words, x is just a value responding to y without an actual meaning. When memorization process happens, link coefficient value increases along y-axis. However when forgetting process happens, link coefficient between words (y) will be transformed into x at first and then x declines by a fixed step, and next computing the new link coefficient according to the declined x. Memory curve is shown in fig 3.

In this figure, value in y-axis denotes coefficient between two words, and the nearer it is to 1, the closer two keywords’ relationship.

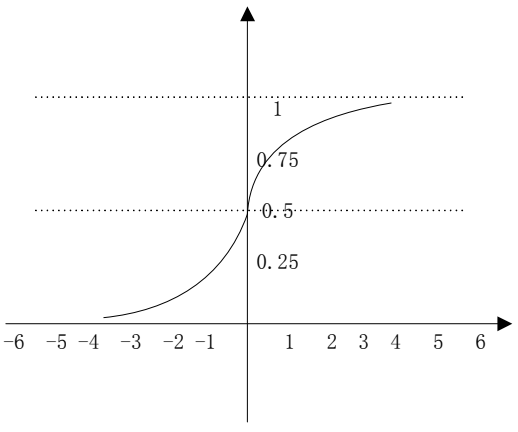

Fig 3.memory curve (sigmoid function)

Definition 1: memorization process

When two words appear concurrently in the same context for the first time, their link coefficient y is set to be a random value, which ranges from 0.1 to 0.2, else y is set according to the following function: $y = y_0 + k_y(1 - y_0)$

y_0 denotes link coefficient before memorization, y denotes link coefficient after memorization process.

(k_y is a memorization coefficient which decides how to strengthen link coefficient between two words ranging from 0 to 1. In this paper, it is set at 0.1)

Definition 2: forgetting process

$$y_0 \xrightarrow{f(y)} x_0 \xrightarrow{x=x0-kx} x \xrightarrow{f(x)} y$$

Firstly, transform y_0 to x_0 according to function x=f(y) , then decrease x_0 according to a fixed step k_x, finally transform x to y according to function y=f(x) and a new link coefficient comes into being.

(k_x is forgetting coefficient which decides how to weaken link coefficient between two words. In this paper, it is set at 0.004)

Definition 3: Filtering

After many iterations of training, some link coefficients between words become too small and they have lost their meaning and we need to get rid of them. In this paper, filtering coefficient is set at 0.2.

Filtering takes place after every 100 documents and in the end of all documents another filtering takes place, which coefficient is 0.3.

Here, an algorithm is proposed to generate a domain word relationship network based on memory principle.

Input: Keywords appearing concurrently in the same document

Key = {key_1,key_2,key_3……key_n}

Output: Domain word relationship network

Step 1:Read a document from the corpus, after segmentation, keywords are available.

step 2:Establish links between each two keywords of this document, add them to the word relationship network, carry out memorization process.

Step 3.Carry out forgetting process to the word relationship network.

Step 4. Read the next documents from the corpus, if it exists turn to Step 2 else turn to step 5

Step 5: Filter word relationship network.

5 Experiments

In this paper, we choose economics part of corpus developed by Fudan University to train the word relationship network[5]. There are 1600 documents all together. Firstly, segmentation based on Apriori takes place, which extracts words occurring frequently in the document, then sorting these words, take the front 5% most frequent words of the whole to be keywords of this document.

Table 1. Part of segmentation

Document id	Keywords of document
1	知识(knowledge) 成本(cost) 开发(develop) 研究(research) 学习(learn) 市场(market) 消费(consume) 跨国公司(multinational company) 知识生产(knowledge based production) 企业(enterprise) 大型企业(big company) 生产(production) 分支机构(branch)
2	消费(consume) 消费需求(consume demand) 收入(income) 经济(economy) 居民(resident) 扩大(enlarge) 住房(house) 低收入(low-income) 提高(elevate) 教育(education) 国家(country) 消费倾向(consumption trend) 发展(development)
3	电子商务(e-commerce) 发展(development) 服务(service) 竞争(competition) 企业(enterprise) 网络(network) 经济(economy) 信息(information) 市场(market) 系统(system) 网上(online) 实现(realize) 政府(government) 消费(consumption) 商家(provider) 银行(bank) 技术(technology) 商务活动(business activity) 美元(dollar) 社会(society) 世界(world) 贸易(trade)
4	企业(enterprise) 产品(product) 市场(market) 知识(knowledge) 消费(consumption) 顾客(consumer) 知识营销(knowledge marketing) 信息(information) 发展(development) 生产(production)
5	贸易(trade) 自由化(liberalism) 中国(China) 进口(import) 出口(export) 模型(model) 关税削减(tariff cut down) 服装鞋帽业(cloth industry) 地区(region) 部门(section) 农业(agriculture) 食品加工业(food industry) 金属矿开采业(mining industry) 综合(overall) 出口需求(export demand) 成员(staff) 变动率(alteration rate) 运输设备业(transport equipment) 中国经济(Chinese economy)
6	财政政策(finance policy) 投资(investment) 国债(national debt) 发展(development) 发行(issuance) 财政(finance) 扩大(enlarge) 市场(market) 积极财政政策(positive finance policy) 经济(economy) 增长(rise) 中国(China) 企业(enterprise) 项目(project) 效益(benefit)

After segmentation of one document, we extract some keywords of the document and establish links between each two keywords and add them to the word relationship network, if two keywords have already linked together in the network then memorization process is carried out to strengthen the link, otherwise, set their link coefficient a random value which ranges from 0.1 to 0.2. When finishing adding all the keywords links, forgetting process will be carried out to weaken all links between keywords in the network. Reading next document and carrying out the same process until all documents are processed. Finally, filtering the links with a threshold, which is set at 0.3 in this paper. A word relationship network with rather high quality is constructed.

A part of the word relationship network is shown as the below Fig. 4:

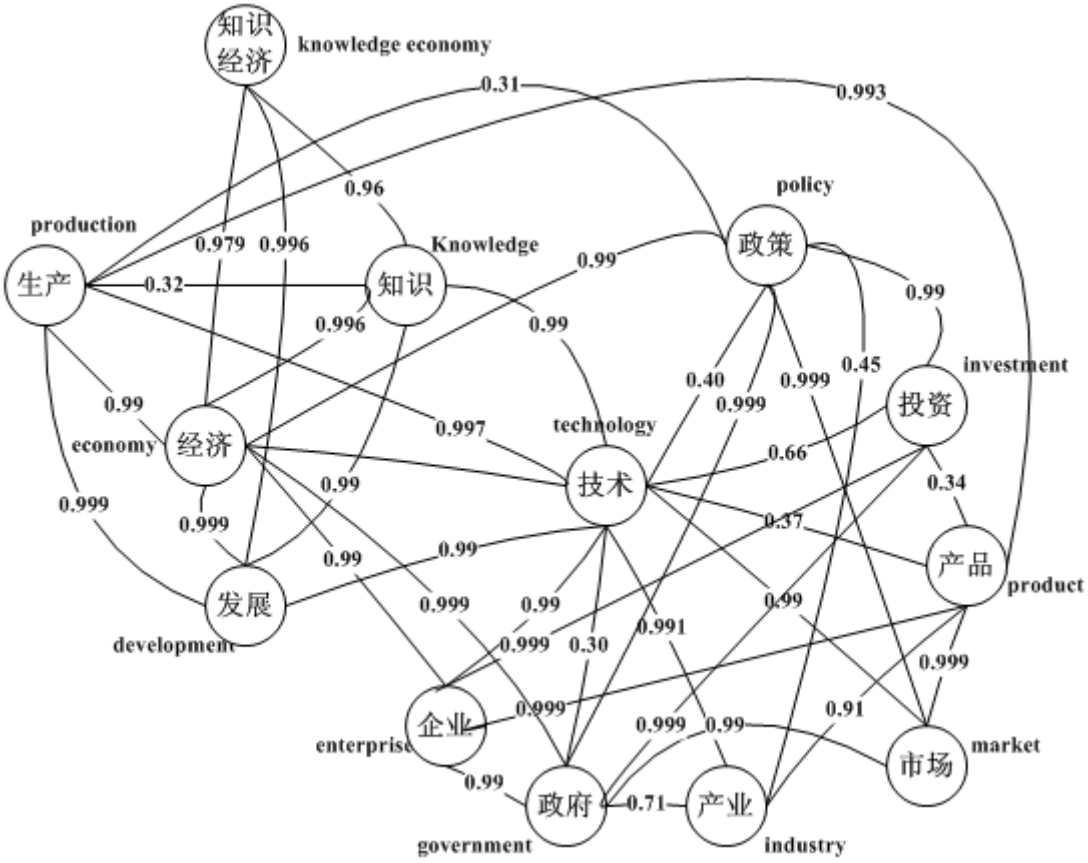

Fig. 4 a part of the word relationship network

In the above figure, a circle denotes keywords and a line denotes link between two keywords on which their link coefficient is labeled. Link coefficient ranges from 0 to 1 and the nearer the link coefficient is to 1, the closer their relationship is.

After training, the word relationship network has filtered lots of non-domain words successfully. The survivals have high quality and reliability. Through training, we extract domain words and their relationship hiding in documents and the last result is more objective.

6 conclusion and future directions

Domain word relationship network created by memory model has laid a signification foundation for further research in many areas,

for example automatic ontology creation, labels for text clustering, semantic analysis of text, topic recognition, and dimension reduction and so on. It will reveal its advantages in dealing with text due to its inherent properties.

In addition, through analysis of the experiment, we found that there are still some aspects to be improved and worthy of further research. Firstly, in memory model, there are three coefficients (memorization coefficient, forgetting coefficient and filtering coefficient) need to be adjusted according to some experiences. Secondly, segmentation quality affects the result of word relationship network. Thirdly, segmentation based on Apriori has a good quality but it is a bit slow, the efficiency of algorithm should be improved further. However, memory model has rather effective function to filter non-domain words.

In segmentation part, stop words are from information retrieval lab of Harbin Institute of Technology [6]. Further research includes text semantic analysis and clustering based on domain word relationship network.

Acknowledgement

The authors are grateful to the editors, referees and the National Natural Science Funds of China (70431001).

References

[1]. Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. WIDM'05, November 5, 2005

[2]. Collin F. Baker, Charles J. Fillmore, John B. Lowe. The Berkeley FrameNet Project, In Proceedings of the. COLING-ACL, 1998

[3]. Chen Tao, Sun Maosong. Automated Construction of Chinese Thesaurus Based on Self-Organizing Map. Journal of the China Society for Scientific and Technical Information. 26 (1): 77-83, 2007.

[4]. Wang Jinmei. On Memory Rule and Learning Strategy of English Words. Journal of ChengDu University of Technology.(Social Science), 14(3): 68-71, 2006.

[5]. Fudan University Corpus. http://www.nlp.org.cn/project/project.php?proj_id=6

[6]. HIT IR-lab stop word. http://bbs.ir-lab.org/cgi-bin/topic.cgi?forum=19&topic=86&show=0