Intelligent Automation And Soft Computing, 2018

Vol.24,no.3,653–661

Association Link Network based Concept Learning in Patent Corpus

Wei Qin, Xiangfeng Luo

School of Computer Engineering and Science, Shanghai University, China

ABSTRACT

Concept learning has attracted considerable attention as a means to tackle problems of representation and learning corpus knowledge. In this paper, we investigate a challenging problem to automatically construct a patent concept learning model. Our model consists of two main processes; which is the acquisition of the initial concept graph and refined process for the initial concept graph. The learning algorithm of a patent concept graph is designed based on the Association Link Network (ALN). A concept is usually described by multiple documents utilizing ALN here in concept learning. We propose a mixture-ALN, which add links between documents and the lexical level, compared with the ALN. Then, a heuristic algorithm is proposed to refine the concept graph, leading to a more concise and simpler knowledge for the concept. The heuristic algorithm consists of four phases; first, for simplifying bag of words for concept in patent corpus, we start to select a core node from the initial concept graph. Second, for learning the association rule for the concept, we searched important association rules around the core node in our rules collection. Third, to ensure coherent semantics of the concept, we selected corresponding documents based on the selected association rules and words. Finally, for enriching semantics of the refined concept, we iteratively selected core nodes based on the corresponding documents and restarted our heuristic algorithm. In the experiments, our model shows effectiveness and improvements in prediction accuracy in the retrieve task of the patent.

KEY WORDS: Association link network, Concept learning, Heuristic algorithm, Patent corpus, Retrieve task

1 INTRODUCTION

WITH the explosion increase of Internet documents, millions of new concepts appear on the Internet as documents and provides a good chance for web user to learn knowledge on it. Every day millions of people search for information through computer, mobile phone and etc. by search engines such as PatentList, Baidu, and Google. In the standard IR algorithm, the vector space mode (VSM) is used to represent user query and documents, and some predefined score function is designed for selecting relevant documents based on similarity between query and document, which is considered as a concept refined method based on user query. However, the standard IR algorithm, which is a precision-oriented task, is where a user finds an answer to their information needs that can typically be addressed by one or two relevant concepts. Such methods cannot satisfy a user's requirement, when the user is typically ready to check possibly hundreds of relevant documents to summarize the concept. For example; the user wants to know "automobile engine" in the patent corpus. The standard IR algorithm may only provide one or two aspects about the automobile engine such as engine power or engine fuel consumption. In this paper, we search an algorithm to construct and refine concepts, which can learn the knowledge in the patent corpus, which can better represent a patent concept. The refined concept is represented as a smaller set of a document list and several core keywords and association rules.

Existing methods of the relevant concept of the learning method such as Hjorland (2009), and the concept theory and learning method by Hammer, et al. (2009), have been widely used to represent specific corpus or domain. However, they have encountered the following challenges: 1) How to construct a concept representation that could be understand both by machine and user, 2) How to construct a user readable concept graph automatically without assistance from domain experts, 3) A number of state-of-the-art

CONTACT Wei Qin qinweidoctor@shu.edu.cn

concept learning methods relying on specific domain and model which transfer another domain, 4) Learning concept approaches, which were adopted by various literature will take exponential growth of time with the increase number of documents.

These problems appear more serious in a patent corpus than a normal corpus such as a new article. A patent is hard to understand, because it exists of large numbers of terminology, which is not well understood both with machine and user. The technology used in a patent is broad, even some has cross-domain technology that requires a multi-domain expert to extract the concept from it.

In this paper, we address a concept learning task as a construct for the initial concept graph and refine the concept problem, which obtains sever documents and word document rules to represent the concept. In order to rely on the above four problems as mentioned, first, we use an extension form of ALN, which we call it MALN, that makes it more convenient for constructing a scoring function. Second, a heuristic algorithm is proposed based on the MALN that consist of four phases: 1) Select Core Nodes Phase 2) Select Association Rules Collection Phase 3) Select Document Phase and 4) Feedback to the MALN Phase. In the following section all the phase will be described in detail.

To address such challenges in the concept learning, we propose our model and make the following contributions:

1) A small set of documents, which incorporate keywords and association rules is used to represent a specific concept, which both increase readability of people and machine.
2) The process of building the ALN is fully automated without a domain expert assistant, moreover, manual participation is not required for the heuristic process.
3) Our model can transplant to any other domain with few changes to the model, and a training set is unnecessary for our model.

The remainder of this paper is organized as follows: We continue by covering the related work in Section 2. Then, we introduce the basic definition and construction method of the ALN and how we expand to the MALN in Section 3. Section 4 describes procedures of our heuristic algorithm in detail. Section 5 shows the experiment to our concept learning model. Finally, we conclude in Section 6.

2 RELATED WORK

2.1 Concept Represent Model

DIFFERENT methods on concept learning have been proposed in literature, and can be divided into three categories; expert based methods, statistics machine learning methods, and user memory based methods.

The expert based methods (Aizawa, 2003), (Guthrie, et. al., 2006) required experts to be familiar with a specific domain that contains concepts the people want to know. (Angluin, 1988) attempted using queries to learn an unknown concept. Several types of queries are used and studied in the supervised learning framework. Jong, K. A. D. (1975) explored the use of genetic algorithms as a key process in the implementation of the concept learning system. He assumes that conceptual learning should be aggregated, which implies that people always focus on several core concept nodes in the process of learning. Some people consider that concept learning should be based on ontology rules (Rouder & Ratcliff, 2006)), which is designed by an expert. In this method, the learning direction decisions are made simply and rely on simple relevant rules without considering rules that exist in a rules set. These concept learning methods need to be a manual participant or at least a semi manual designation of some feature, which are limited to the particular field. The result of learning is only the relationship among the features or the features and concepts, which greatly increases the people's cognitive burden.

The statistics machine learning methods model concept is by one or more vectors. The vector space mode (VSM) (Salton, 1971)) has been successfully applied to the famous SMART document retrieval system. Contents of the document are reduced to vector space operations, and it uses spatial similarity to express semantic similarity, is intuitive and easy to understand. PLSA (Kushilevitz, et al., 1998) is an index retrieval method. The method and the traditional vector space model (VSM) used as a vector to represent the word (terms) and document(s), and the relationship between vectors (such as angle) to determine the relationship between words and documents. Exception is, LSA, which will be mapped to the document words and latent semantic space, which in addition to some of the "noise" of original in the vector space, improves the precision of information retrieval. Word2vec (Mikolov, et. al., 2013) is a distributed representation other than One-hot Representation in the traditional concept representation such as VSM mentioned above. It grants certain semantic meaning to words, so that we could easily calculate the similarity between two vectors. Although this type of approach could be easily handled by a machine, it lacks an intuitive understanding of the concept. Simultaneously, some of these methods only have an initial representation for the concept such as VSM; others need training for vectors and are hard to transfer to another domain such as Word2vec.

Some research represents a graph-based concept, which is a more interpretable concept of the representation that can be understood by both user and machine (Hussain, et al. 2014). ALN (Luo, et al., 2011) is one of a typical user memory based method for

concept representation, which is more perceptual intuition. ALN could not only be defined by formal but, also can convert into a concept graph naturally. It is efficient and effective for the construct concept, and the graph structure makes it easier to interfere in the following task. We consider ALN as our initial concept representation, and extend it to MALN for applying to the heuristic algorithm.

2.2 Evaluation of Concept Learning

At present, the evaluation of concept learning has not formed unified evaluation criteria. Researchers usually use artificial methods to evaluate the quality of construction concepts, or evaluate them through conceptual applications. In this paper, we focus on the IR method, which is one of most popular application based on construction concepts. IR's query could be considered as part terminology, which is augmented by finding a relevant document. Some research already uses an image retrieval task to evaluate the result of an image concept domain (Huang, Hu, et. al., 2016). It is common practice to use heuristic rules to construct a document ranking list in a search. Usually, rules are created based on the observation about the relationship between query and document. Most of a heuristic is designed for advanced similarity scoring function between query and document in a corpus. The most known framework is OKAPI (Robertson& Zaragoza, 2009), the Scoring function fused word frequency, and the document length is featured into one function. Subsequently, Query reformulation (Mahdabi & Crestani, 2014)) method was proposed to avoid noise existence in the query of IR task, which successfully is applied to news articles and patent prior art search, another study, using query expansion via feature selection (Zhang, et. al., 2016). Both the query formulation method and its expansion have shown better performance compared to using the query as a core concept. However, the refined method of IR cannot learn a complete concept representation when the user query is incomplete, which frequently occurs. And, when the only document selected by the IR method cannot represent the true nature of the concept. Therefore, we propose a new heuristic algorithm applied to select documents, keywords, and rules about the concept.

3 BUILDING MIXTURE ASSOCIATION SEMANTIC LINK NETWORK FOR THE CONTENT REPRESENTATION OF CONCEPT

3.1 Frame Work

IN this paper, we propose a model for automatic concept learning that has good representation and quality from the patents data. We use ALN to construct and represent the concept that people desire. Then we generate a summarization from the given paper by using a network last phase. The two phases we will describe subsequently. The whole process of our model is shown in Figure 1:

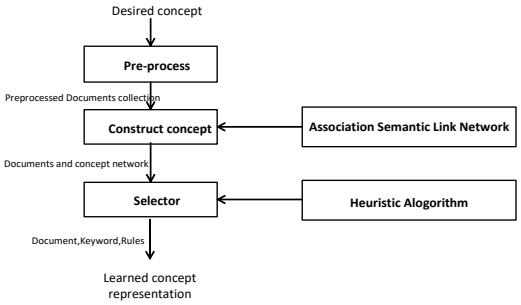

Figure 1. Concept Learning Model Framework.

The pre-process is uses standard Natural Language Processing steps, which reduced the document noise in the patent corpus, and normalized the document, we will discuss it in our experiment. The Constructed concept is not only of a concept representation, but also is the foundation for the subsequent use of the heuristic algorithm for learning concepts. This step will be illustrated later on. The selected document obtained by a heuristic algorithm is the core of our model, which selects a document, rules and keywords that could properly represent the concept. It is able to learn concepts effectively from the constructed concept graph. This step discussed in Section 4.

3.2 Building an Association Semantic Link Network

The traditional concept of construction and establishment of may requirement for artificial participation, and some even completely rely on relevant field experts. Therefore, although the accuracy of conventional methods are relatively high and easy to understand, it is difficult to handle growing number of documents in the Internet, especially new concepts.

The Association Semantic Link Network (ALN), (Luo, et al., 2011) is a resource organization model, which is used for extracting a core sematic and store correspondence knowledge. Given a document list comes from a query, ALN automatically learns the document list representation based on the co-occurrence information of a word and topological of the graph structure. The key principle of the ALN is the document list representation be converted into a word-based graph. The ALN is defined in (1):

$$ALN = (N, L) \tag{1}$$

where $N = \{n_i \mid 1 \le i \le n\}$ is the corresponding vector represented document, and keywords are extracted from the document. The weight associated with the keyword is calculated by TF-IDF, where TF is the frequency of the keyword in the document and IDF is the inverse document frequency of the keyword. The

length of N indicates the number of keywords that exist in the document. $L=[L_{ij}]_{n\times n}$ is a rule matrix is defined in (2):

$$L_{ij}=\frac{\#(\mathrm{i,j})}{\#(i)\square\#(j)} \tag{2}$$

Here #(i,j) is the number of times word j appears in the context of word i , #(i) is i IDF frequency #(j) is j IDF frequency.

After making the connection using formula 1, ALN optimized itself the structure according to the small world (Collins, & Chow, 1998) and the scale free network theory (Yoo, & Hu, 2006). As ALN represents the semantic association and has the ability of extracting the core semantic from the document, which is consistent with the concept definition. We have sufficient reasons to use production of the ALN as an initial representation of the concept.

3.3 Building a Mixture Association Link Network

The simplest form of the ALN only represents graphs of lexical layers, however for a document list, it is necessary to represent both lexical layers and document layers. Therefore, we slightly extended the definition of the original ALN, so that it has stronger ability to express the relationship between the documents and the lexical, it is a Mixture Association Link Network (MALN), which is defined in (3):

$$MALN=(S,N,L,D) \tag{3}$$

where the definition of N, L is the same as formula 1. $S=\{s_i\,|\,1\le i\le s\}$ is the corresponding vector of the document that is related to the concept. We consider each document having same importance so that the vector weight set is 1. Given the document collection S and association rules collection L, we have a dependent relationship from S to L, which is defined as $D=[D_{ij}]_{s\times n^2}$ and $D_{ij}=1$ indicates that document i has rule j (Goodman, et al. 2010).

According to the definition, MALN can be converted to a multi-level graph, which the upper layer represents S, the bottom layer represents N. L, and D is the Links for the two layers. The graph is a foundation used for the heuristic algorithm. Figure 2 gives the graph illustration of the MALN.

4 HEURISTIC ALGORITHM FOR A CONCEPT SELECTOR

OUR algorithm first started from several core keywords as a starting point for MALN. Secondly, we find the best collection of association rules for the moment through the breadth traversal of graphs. Third, we select the most relevant document to become active

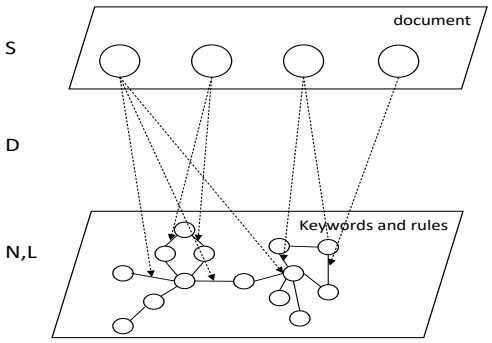

Figure 2. Illustration of MALN

according to the association rules collection. Fourth, the selected document gives feedback to the MALN and returns to phase2. The algorithm stopped when the core keywords and association rules are selected.

In Fact, our heuristic algorithm for concept learning is simulated from the real user behavior describing the concept, and it is a clear oriented algorithm that is based on specific concepts, which a user wants to know. When a user wants a concept, he releases some obscure keywords, which always is the most relevant terminology about a concept. Then, he will search relevant documents based on these core keywords to know more about the knowledge concept, which expands the concept terminology and construct Link between the knowledge of the terminology. Finally, the user will have a good grasp of the concept by repeating the above mentioned two phase.

In this section, we describe the proposed heuristic algorithm. We start by describing the procedure of the four phases in our algorithm, and then we describe our approach procedure with an algorithm depicted.

4.1 Select Initial Core Nodes Phase

The meaning of the select core node is a common concept that can always be of one or more words to represent, where the core node comes from two main approaches. As our heuristic method is an iterative algorithm, second sources of the core nodes are only calculated when having an initial document list, which we will discuss in phase 4. At the beginning of heuristic algorithm, we accept two types of initial core nodes, which come from user specified keywords such as; query word or comes from a category description word; however we discover that the experiment received poor performance if we put all query words into the core nodes. Therefore, we designed a pre-defined formula (4) to remove noisy words existing in specified keywords:

$$\mathrm{CN}=\{\mathrm{w}\in Q\,|\,DF<\alpha^1\} \tag{4}$$

Here, we calculate document frequency (DF) for every word in a specified keywords list, notice that we consider the query word type as same as a category

description word. We removed words with higher average frequency and added all the rest of the words to the core node list. We show a graph depict of a selected core nodes phase in Figure 3:

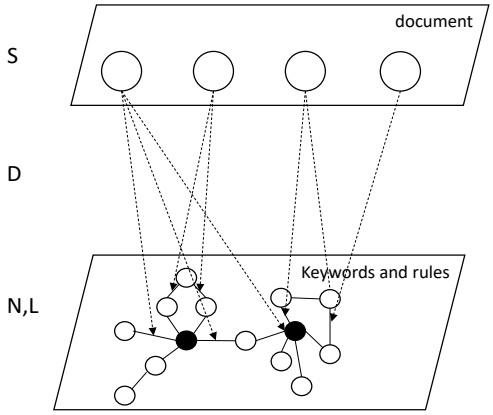

Figure 3. Select Initial Core Nodes.

4.2 Select Association Rules Collection Phase

After selecting the core node form MALN, we will select rules around the core node, which we assume the user will search for relevant concepts given certain core nodes. The associate rules are stored in L, so our task is to select some maximizing weight sub-rules collection based on the core node, the rule weight is calculate by formula (5).

$$RC_{ij}=\{rule \in L_{ij} \mid \frac{rule(i,j)}{\sum_{(i,j)\in S} rule(i,j)} > \beta\} \tag{5}$$

$$i \text{ or } j \in \{\text{selected core node from phase 1}\}$$

Where L_{ij} is representing of the definition of MALN and β is the threshold for selecting the Association Rules Collection. Consider the core nodes that are extracted from last phase as a starting point. We find the maximum weight association rules from surround that satisfy at least one of constraints add to association rules collection every time. The constraints include: 1) One end of rules must be an initial keyword node, because our rule collection phase aims to find the most relevant rules for the concept. 2) When the weight of the rule is the same, we prioritize select rules that where both ends nodes are in the core node, which have a great possibility of becoming an exclusive phrase in the concept. The constraint above mentioned is to ensure that the association rules could be a more complete representation for the concept. Figure 4 shows the result of a select association rules collection phase.

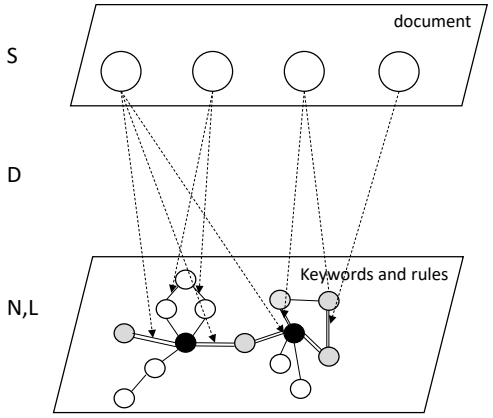

Figure 4. Select Association Rules Collection.

4.3 Select Document Phase

Different from the IR method, we select relevant documents based on both selected core nodes and a selected rule collection. The relation tuple D in MLAN, which is standard for which rules and keywords are contained in document D and are used to calculate the Score for each document in a corpus. We sum the weight of the selected core nodes and rule collection in the document, the highest score of the document is selected to our model: Scoring function is defined in formula (6):

$$Score(d)=\frac{\sum_{rule\in RC \wedge rule\in d} weight_{rule}+\sum_{w\in UK \wedge rule\in d} weight_{w}}{\sum_{rule\in d} weight_{rule}+\sum_{w\in d} weight_{w}} \tag{6}$$

Where the $weight_{rule}$ is the rule's weight calculated in MALN and $weight_{w}$ is the word's weight in MALN. We normalize the document score by the sum of all rules weight and word score. Then several maximum score documents form Score(d) will be put into our model as one representation for our concept representation. Figure 5 shows the result of select document phase.

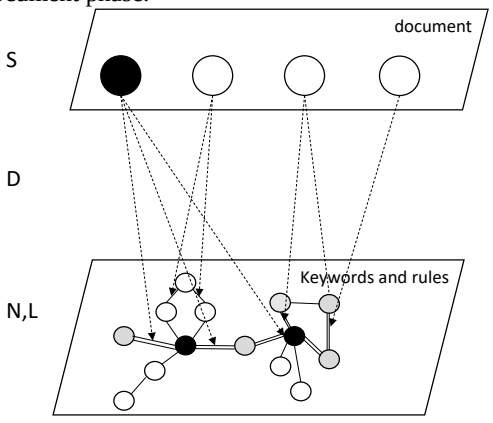

Figure 5. Select Document.

4.4 Feedback on MALN Phase

After we selected one or more documents into the concept representation, we will change the core nodes list and rules list and call it the feedback phase. In this phase, we focus on the problem of the change core nodes, in other words, which core node will be activated on the MALN. We calculate the words TFIDF value as a score function where a super threshold is selected into the core node. Then, the rules are reselected in the Select Rules Collection Phase. The words TFIDF function is defined as follows:

$$CN_w = \{w \in Q \bigcup D \mid \sum_{w \in Q \bigcup D} freq(w) * IDF(w) > \alpha^2\} \quad (7)$$

Here, w is one of the words that exist in a query concept words list or in a selected document. It is worth noting that we sum all frequencies of w in the selected document as TF, instead of only occurring in an individual document. In addition, the $IDF(w)$ value is calculated in all the documents. Formula 6 indicates that we prefer selecting words that frequently occur in a document obtained from the previous phase, and prefer not to select a common sense word. Figure 6 shows the result of the feedback on the MALN phase.

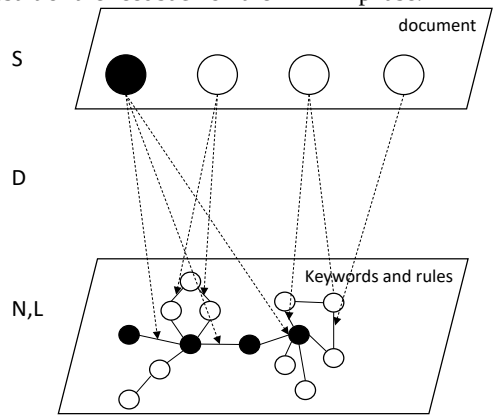

Figure 6. Feedback on MALN

We program our algorithm as follow:

Algorithm 1. Heuristic algorithm to refine concept

Input: MALN(S, N, L, D)
User Keywords: UK
Core Nodes: CN
Rules Collection: RC
Core Documents:CD
Output: Concept Represent ={core documents, keyword, document rule}

Initialize Concept Represent = ϕ
CN=UK //Select Core Nodes
While |SUMM|<LIMIITED
if iteration=1
CN=UK
for (rules around the {CN})
If RC_{ij} > threshold then // RC_{ij} defined in formula 5
RC+= RC_{ij}
end for
CD+= argmax(Score(d))// Score(d) defined in formula 6
for (nodes in {CD})
If $CN(w_k)$ > threshold then // $CN(w_k)$ defined in formula 7
CN += CN_w
end for
end while
return Concept Represent

5 Experiments and Analysis

5.1 Dataset and Pre-process

WE conduct the experiment on the china patent document dataset from the State Patent Bureau. The China patent document consist of a title, abstract, claim, description, applicant, inventor, publication date and International Patent Classification codes (IPC).We utilize the IPC description as our initial core nodes in our model, and applied it to the classification task, which is the first upset order of the IPC document, then we redistribute the unordered the document to correspond to the IPC code. Our concept learning model is applied to the learn patent concept. We select four IPC categories from the third-level IPC, and the description of the corresponding IPC are; how to dry objects, deposit box, steam, toy car, own number of (356, 356, 357), 357 documents respectively. The goal of our model is make people learn these four concepts fast and effectively. We used a Stanford-segmenter to segment the Chinese content. Then the stop words that were provided by the Harbin University and BaiDu's stop word list are removed from the document.

5.2 Evaluation Strategy of our Concept Learning Model

We evaluate the quality of our concept representation by performing a retrieval task. The former will evaluate a concept, which represents potential effectiveness in predicting if the other relevant documents belong to a specific concept. Specifically, the selected documents and keywords, and rules in the MALN from our leaning model will be considered as the retrieval centre of the concept. Once we have identified the retrieval centre, the rest of documents could be directly assigned to a corresponding Category based on some distance formula. The cos similarity, which is defined in formula (8) is used in our experiment as the distance formula:

$$sim(c_i, d_j) = \frac{c_i \cdot d_j}{\| c_i \| \times \| d_j \|} \quad (8)$$

Here, d_j document is a vector in D, which the weight is represented as TFIDF. c_i is the centre of the concept, which is a vector consisting of a selected document, keyword, and rules extracted in our model. After all the documents in corpus are classified, we use a precision score criteria, which is commonly used in retrieving and clustering a task to evaluate our result.

Although we evaluate our concept learning model only using documents information, representation of the keywords and rules are more important for the concept. However, it is difficult to evaluate on just keywords and rules layer in the patent corpus. There is not an existing automatic method to directly evaluate two representations in a patent apart from human assistance as we know.

5.3 Results of the Concept Learning Model Evaluation

We take four classes extracted before as our test data sets. We need to learn the four independent concepts from the four classes. To address how we generate the selected concept representation from our model, we briefly describe the process of the learning concepts from our data sets. First, we put four categories of the corpus of documents together as our concept learning data without its category label. Second, based on the mixed corpus, MALN is constructed as an initial representation of the documents data. Third, we select a category description such as toy and car as our core node in the heuristic algorithm. Subsequently, the completed heuristic algorithm is applied for a select document, keywords and rules.

Through a great deal of the experiment, we found that of the hyper parameter, the β defined in phase 2 in the heuristic algorithm will have greater impact on our model. We empirically seek to evaluate the threshold β defined phase two in the heuristic algorithm yielding the best performance. Parameter β varies from 0 to 1 with 0.2 intervals. In this experiment, we run 10 times our heuristic algorithm to select 10 documents and corresponding keywords, and rules as learned knowledge. When β is 0, it represents our rule collection only having one rule, while results to a variety of documents will have the same score with a selected document. When β is 1, the algorithm will degenerate to a global rule select function, which almost neglects a related constraint in formula (4). We chose β as 0.2, 0.4, 0.6, 0.8, 1, 0 is excluded from our test data, because, the document selection phase becomes a random selection process. Figure 7 shows the precision score of our corpus from four categories retrieved as a qualitative is evaluated.

Figure 7. Precision vs Threshold β for Retrieving Task.

Notably, there is a rather steep drop-off in performance when β >0.4, that is, too many irrelevant rules are included in the association rule collection. Therefore, the parameter of β =0.4 is set to the best threshold in our experiments.

Figure 8 reports the performances of the proposed methods and baseline BM25 in the IR system. We use the category description as a query to the BM25 and extract the most relevant document as a core concept representation. Then, a simple classifying method is proposed in evaluating the strategy and is applied to evaluate the precision score. In order to compare the two methods in a similar environment, the BM25 selected document list size is the same as our model.

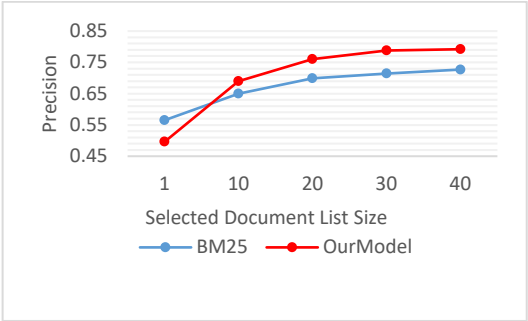

Figure 8.Models Performance vs. a Selected Document List Size for Two Selected Approaches.

The results indicate that improvements of our model over the baseline BM25 are significant. For example; compared to the best result when the selected document list size is 20, our model improves 7%. The results also indicate that 30 documents are enough to represent the patent concept, which testifies the assumption of every concept that could be represented by a limited number of document, keywords and rules.

6 CONCLUSION

ACCURACY, conciseness and comprehensiveness are three important criteria for evaluating the constructed concept in a patent corpus. In this paper, we proposed an automatic concept learning model, a mixture ALN graph for running the heuristic algorithm

for a patent concept learning, to improve accuracy, conciseness and comprehensiveness of patent concepts.

1) The mixture ALN, which is expanded from the ALN via the add relation rules between the document and lexical, are used to express the initial patent concept. This approach improves the comprehensiveness of the concept. And such knowledge representation of the concept can be expressed in graphs, which enables us to use a graph based heuristic algorithm.
2) The select core node phase aims to improve the concept accuracy, which focuses on the extract user designated concept keyword and concept terminology from a corresponding document. This approach removes a noise word, which improves concept accuracy.
3) The select association rule collection phase, which finds the most relevant association rules around the concept terminology. This approach will easily find additional terminology about concept so that improving both concept conciseness and comprehensiveness.
4) The select document phase and feedback phase in our heuristic algorithm, which can acquire relevant documents and reselect the core node and rules collection based on the document. This approach purifies the noise caused by the association rule expansion, while avoiding the overestimation of the irrelevant concept terminology

7 ACKNOWLEDGMENT

THE research work reported in this paper was supported in part by the National Science Foundation of China under grant nos. 61471232. This work was jointly supported by the Shanghai Science International Cooperation Project under grant no.16550720400

8 REFERENCE

A. Aizawa. (2003). Title, *An information-theoretic perspective of tf–idf measures .Information Processing & Management*, 39(1), 45-65.

D. Angluin. (1988). Title, *Queries and concept learning. Machine Learning*, 2(4), 319-342.

J. J. Collins & C. C. Chow. (1998). Title, *It's a small world. Nature*, 393(6684), 409.

N. D. Goodman, J. B. Tenenbaum, J. Feldman, & T. L. Griffiths. (2010). Title, *A rational analysis of rule-based concept learning. Cognitive Science*, 32(1), 108-154.

D. Guthrie, B. Allison, W. Liu, L. Guthrie, & Y. Wilks. (2006). Title, *A Closer Look at Skip-gram Modelling*. (pp.1222--1225).

R. Hammer, G. Diesendruck, D. Weinshall, & S. Hochstein. (2009). Title, *The development of category learning strategies: what makes the difference?. Cognition*, 112(1), 105-119.

B. Hjorland. (2009). Title, *Concept theory. Journal of the American Society for Information Science & Technology*, 60(8), 1519-1536.

T. Huang, X. Hu, & S. X. Yang. (2016). Title, *Networks based computing and automation*, 22(4), 533-534.

S. S. Hussain, M. Hashmani, M. Moinuddin, & K. Raza. (2014). Title, *A novel topology in modular ann approach for multi-modal concept identification and image retrieval. Intelligent Automation & Soft Computing*, 20(1), 131-141.

K. A. D. Jong. (1975). Title, *Analysis of the behavior of a class of genetic adaptive systems. Ph.D. thesis University of Michigan.*

E. Kushilevitz, R. Ostrovsky, & T. Rabani. (1998). Title, *Efficient search for approximate nearest neighbor in high dimensional spaces. Thirtieth ACM Symposium on Theory of Computing* (Vol.30, pp.614-623).

X. Luo, Z. Xu, J. Yu, & X. Chen. (2011). Title, *Building association link network for semantic link on web resources. IEEE Transactions on Automation Science & Engineering*, 8(3), 482-494.

P. Mahdabi & F. Crestani. (2014). Title, *Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence*. ACM.

T. Mikolov, K. Chen, G. Corrado, & J. Dean. (2013). Title, *Efficient estimation of word representations in vector space. Computer Science.*

S. Robertson & H. Zaragoza. (2009). Title, *The probabilistic relevance framework: bm25 and beyond. Foundations & Trends® in Information Retrieval*, 3(4), 333-389.

J. N. Rouder & R. Ratcliff. (2006). Title, *Comparing exemplar- and rule-based theories of categorization. Current Directions in Psychological Science*, 15(1), 9-13.

G. Salton. (1971). Title, *Experiments in Automatic Thesaurus Construction for Information Retrieval. In Proceedings Ifip Congress, Ta-2* (Vol.71, pp.115-123).

I. Yoo & X. Hu. (2006). Title, *Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory. International IEEE Conference on Intelligent Systems* (pp.851-858). IEEE.

Z. Zhang, Q. Wang, L. Si, & J. Gao. (2016). Title, *Learning for efficient supervised query expansion via two-stage feature selection.* 265-274.

9 NOTES ON CONTRIBUTORS

Wei Qin received bachelor's degree in 2013 from XiDian University, China. Currently, he is a student doctor at Shanghai University. His main research interests include concept learning and knowledge graph

Xiangfeng Luo is a professor in the School of Computers, Shanghai University, China. Currently, he is a visiting professor at Purdue University. He received master's and Ph.D. degrees from Hefei University of Technology in 2000 and 2003, respectively. He was a postdoctoral researcher with the China Knowledge Grid Research Group, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), from 2003 to 2005. His main research interests include web wisdom, cognitive informatics, and text understanding. He has authored or co-authored more than 50 publications and his publications have appeared in IEEE Trans. on Automation Science and Engineering, IEEE Trans. on Systems, Man, and Cybernetics-Part C, IEEE Trans. on Learning Technology, Concurrency and Computation: Practice and Experience, and New Generation Computing, etc. He has served as the Guest Editor of ACM Transactions on Intelligent Systems and Technology. Dr. Luo has also served on committees of a number of conferences/workshops, including Program Co-chair of ICWL 2010 (Shanghai), WISM 2012 (Chengdu), CTUW2011 (Sydney) and more than 40 PC members of conferences and workshops.