Knowledge Annotation and Discovery for Patent Analysis

Su-Hsien Huang
Department of Computer and Information Science, National Chiao-Tung University
sshuang@cis.nctu.edu.tw

Cheng-Chang Liu
Department of Computer and Information Science, National Chiao-Tung University
gis92573@lib.nctu.edu.tw

Chun-Weng Wang
Department of Computer and Information Science, National Chiao-Tung University
gis92616@cis.nctu.edu.tw

Hao-Ren Ke
Library of Naltional Chiao-Tung University
claven@lib.nctu.edu.tw

Wei-Pang Yang
Department of Computer and Information Science, National Chiao-Tung University
wpyang@cis.nctu.edu.tw

***Abstract**-Nowadays, techniques of information retrieval and nature language processing have been gradually employed to achieve the task of patent processing. Therefore, employing these techniques advance the knowledge annotation and discovery for patent documents to be more accurate and convenient. This paper develops a patent retrieval system and aims at introducing new algorithms to provide high-precision patent analysis services, including the development of core techniques and the construction of domain knowledge. The syntactic and semantic analysis of patent documents also applies on patent document retrieval to provide advanced patent services. The achievement of this paper contains the following parts: 1) the online patent search, 2) the structure clustering of patents, 3) the patent summarization and 4) the patent trend tracking.*

Keywords: Patent Search, Ontology, Structured Clustering, Topic Tracking, Summarization.

1. Introduction

Intellectual property (IP) content is quite important knowledge of human being. For enterprises, much attention is paid to intellectual property for tracing the trend of patent development or deafening IP accusation. However, increasing patent documents make the reading complicate and hence requires advanced information technology to assist the investigation of patents. Nowadays, techniques of information retrieval and nature language processing have been gradually employed to achieve the task of patent corpus processing [1][2]. The analysis of patent documents can be more accurate and convenient through these advanced techniques [3].

This paper aims at the invention of novel algorithms to provide high-precision patent retrieval, including the development of core techniques and the construction of domain knowledge. The syntactic and semantic analysis of patent documents is applied to patent retrieval for providing advanced patent services, include patent clustering, patent summarization and patent trend tracking. A complete structure of patent retrieval system is shown to annotate and discover knowledge by the proposed algorithms from patent documents. The following sections introduce each component and their algorithms.

2. The Patent Retrieval System

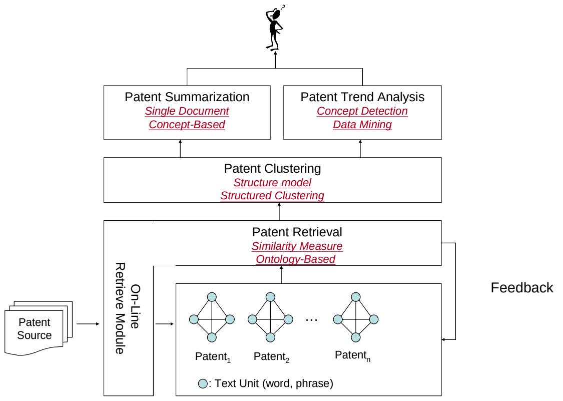


Figure 1. The Patent Retrieval System

The patent retrieval system provides on-line search from some patent sources. The patent sources may be USPTO (http://www.uspto.gov/) or WIPO (http://www.wipo.org), etc. A user submits keywords to the on-line retrieve module for retrieving documents from the patent sources. The retrieved documents are firstly analyzed in the patent database to extract patent concepts and construct domain ontology. In the meanwhile, the user can preview and choose patent concepts to rebuild his query and extend the search scope by domain ontology. The search result is then classified in the patent clustering module. Finally, the clustering results are consequently summarized and analyzed to track the trend. The architecture of the patent retrieval system is depicted in Fig. 1.

The system contains four main parts: 1) the on-line patent search, 2) the structure clustering of patents, 3) the patent summarization, and 4) the patent trend tracking. The whole system is developed in Java SDK 1.4.2. and available in http://www.database.cis.nctu.edu.tw/

2.1. Concept and ontology construction

In our system, retrieved patents are analyzed for extracting connected concept units that are the basis to measure patent similarity. Each unit represents an independent concept. Two or more units are connected when they co-occur frequently. Concept units are segmented from sentences of patents; a single sentence can contain multiple concepts. Additionally, similar concepts are grouped into "concept clusters". The search feedback exploits the clusters to refine user's query.

Ontology is referred to extend the search scope and is beneficial to search. There are two types of ontology in the system:

- Domain knowledge from patents
- External ontology

The first ontology is obtained by calculating the most frequent TF*IDF of uni-, bi- and tri-grams terms from the patents. The relations among concepts in the ontology are determined by domain experts. The second one comes from the existent ontologies. Our system refers to WordNet to extend ontology. For example, "Algorithm", "Algorithmic Rule", "Algorithmic Program" and "Formula" are in the same ontology tree. (Fig. 2)

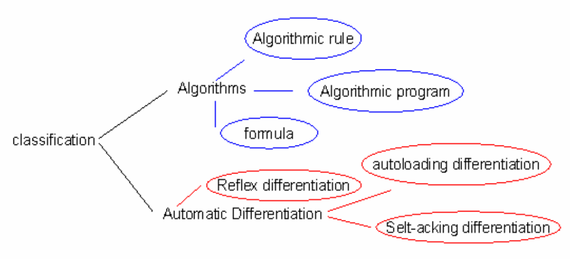


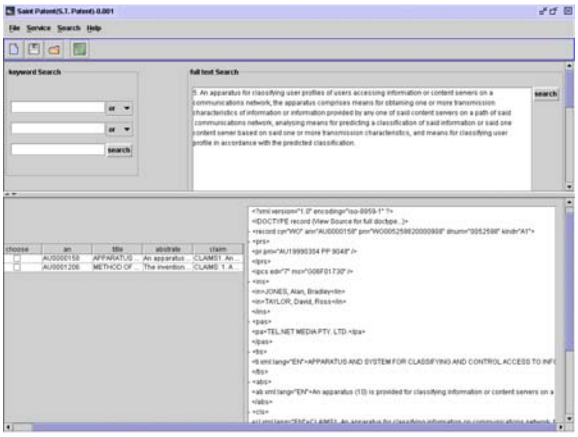
Figure 2. The Ontology tree

2.2. Query expansion

Our system refines user's query by expanding query concept. In the feedback process, the user chooses appropriate concepts from "concept clusters" to compare with connected concept units. The concepts connected with chosen clusters denote related ideas with user's query and can be used to expand the query.

Moreover, user query is also expanded by domain ontology. By comparing with the ontology tree, different words contained in the same ontology group can be identified as related.

Fig. 3 displays the interface of the ontology-based patent search. A user can specify keywords or full-text document to search patents. After analyzing concepts, the search result shows the patent content and concepts extracted from the patent.



5. An apparatus for classifying user profiles of users accessing information or content servers on a communications network, the apparatus comprises means for obtaining one or more transmission characteristics of information or information provided by any one of said content servers on a path of said communications network, analysing means for predicting a classification of said information or said one content server based on said one or more transmission characteristics, and means for classifying user profile in accordance with the predicted classification.

choose	an	title	abstrate	claim
□	AU0000158	APPARATUS ...	An apparatus ...	CLAIMS1. An ...
□	AU0001206	METHOD OF ...	The invention ...	CLAIMS 1. A ...

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<!DOCTYPE record (View Source for full doctype...)>
<record cy="WO" an="AU0000158" pn="WO0052598 20000908" dnum="0052598" kind="A1">
<prs>
<pr prn="AU19990304 PP 9048" />
</prs>
<ipcs ed="7" mc="G06F01730" />
<ins>
<in>JONES, Alan, Bradley</in>
<in>TAYLOR, David, Ross</in>
</ins>
<pas>
<pa>TEL.NET MEDIA PTY. LTD.</pa>
</pas>
<tis>
<ti xml:lang="EN">APPARATUS AND SYSTEM FOR CLASSIFYING AND CONTROL ACCESS TO INF
</tis>
<abs>
<ab xml:lang="EN">An apparatus (10) is provided for classifying information or content servers on a
</abs>
<cls>
<cl xml:lang="EN">CLAIMS1. An apparatus for classifying information on communications network
```

Figure 3. Ontology-based patent search

3. Structured Clustering of Patents

The patents with related concepts are returned in the preceding step. Our system provides a clustering module that is used in the patent summarization and topic tracking service. However, a single patent may contain multiple concepts in its content and hence decrease the performance of clustering. Therefore, the clustering of patents must provide finer-level granularity to improve clustering efficiency.

In this paper, we propose a clustering method considering the structures of patents. Each patent is divided into hierarchical sub-structures (like Claim, Description, paragraphs in a claim, etc.). Concept contains in sub-structures of patents are clustered individually and furthermore the whole patents are clustered according to the sub structure. Clustering according to structure is called structured clustering [6][7][8]. The distinction between structured and conventional clustering is the former provides finer granularity like sentences and paragraphs to obtain better clustering results.

Writing a patent document is required to obey particular convention and style. Therefore, structure in patents can be analyzed heuristically. The analysis of patent structure finds in some literatures [4][5]. In our system, patent structure among sentences is analyzed especially in "Claim" and "Description" sections.

3.1. Structure model

We model a patent document into Directed Acyclic Graph (DAG). Each "node" in DAG represents a paragraph or sentence in the patent and an arrow denotes a structure relationship. Assuming the sequence to fill nodes is breadth first search, the structure of Node S can be represented as follows:

Node_S : ($V_{\mathrm{Node\ S}}$, Node_1, ..., Node_N) ....Eq. 1

where $V_{\mathrm{Node\ S}}$ is the feature vector of Node S. $\mathrm{Node}_{ii\leq N}$ are the coordinate of branches of Node S. N is the maximum number of branches in all nodes.

Fig. 4 illustrates a Chinese patent with analyzed structure. As shown in Fig. 4, if the maximum number of branches is 3, the Node S is represented as Node_s : (V_{Nodes}, Node_1, Node_6, NULL). Node 1 is Node_1 : (V_{Node1}, Node_2, Node_3, Node_5).

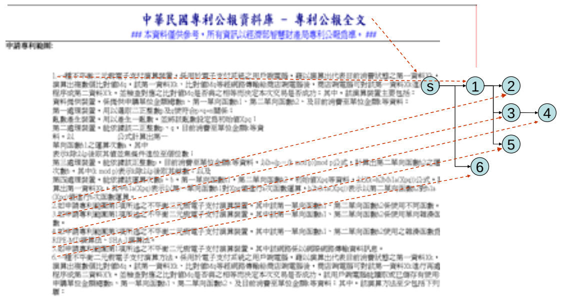


Figure 4. Patent Structure

3.2. Self-organizing map with sructure

The structured clustering in our system refers to Hagenbuchner's Self-Organizing Map (SOM) clustering algorithm in structured data [9]. SOM provides unsupervised neural network clustering and maps high-dimension data into a low-dimension map (usually two). In SOM, closer nodes in the map imply shorter distance in real data. SOM applies in many domains, like bio-structure clustering, graph structure clustering and audio-pattern clustering, etc., and receives good performance. We apply SOM in patent documents clustering with structure considering.

There are five steps to train SOM.

1. Initialize weight vectors of output map as the same number features with input document vector.
2. Present input documents in order.
3. Compute the distance between the input document and all nodes in the map and select the closest node as the winner.
4. Update the weights of the winner node and its neighbors.
5. Repeat step 3-4 to other documents and iterate all inputs until convergence. Label the regions of the final map to represent the clustering result.

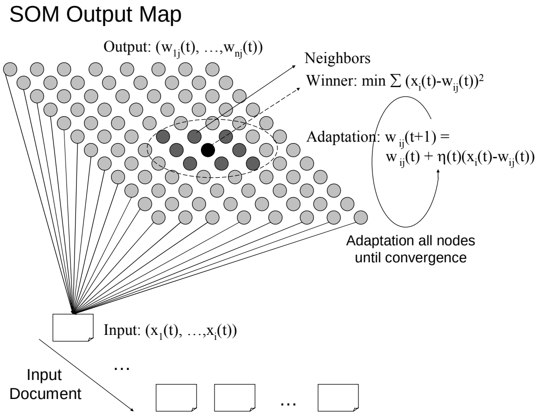


Figure 5. SOM

The training of SOM is illustrated in Fig. 5. Adding structures in SOM requires some changes. Assume the dimension of SOM is 2 and the maximum number of branches is 3. The output nodes are represents as:

$d_{x,y}$ = ($V_{dx,y}$, (x_1 ,y_1), (x_2, y_2), (x_3,y_3)) ………Eq. 2

where (x_i, y_i) is the coordinates of node_i and $V_{dx,y}$ is the feature vector of (x, y). The distance of input and output nodes requires the original distance and the distances of all sub-structure nodes. By referring to Eq. 2 and 3, the distance of the example in Fig. 3 is calculated as follows:

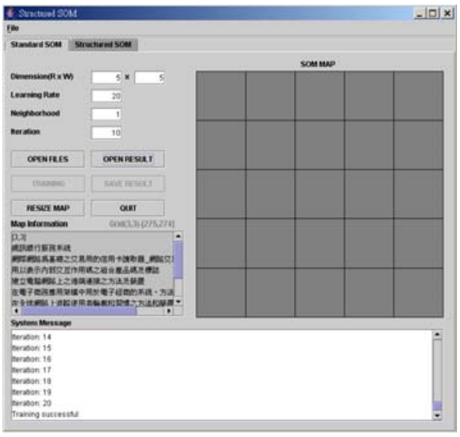
$$d = \sqrt{(V_{Node1} - V_{dx,y})^2} + |V_{Node2} - V_{dx1,y1}| + |V_{Node6} - V_{dx2,y2}| + |NULL - V_{dx3,y3}|$$
……………………………………………………Eq. 3

where $|V_{\mathrm{Node}i}-V_{dxj,yj}|$ is the distance between input Node i and output node (x, y). The adaptation of structure nodes needs to update root and all sub-structure nodes. The formula is shown in Eq. 4:

$w_{dx,y}(t+1) = w_{dx,y}(t) + \eta(t)*|V_{\mathrm{Node1}}-V_{dx,y}|$
$w_{dx1,y1}(t+1) = w_{dx,y}(t) + \eta(t)*|V_{\mathrm{Node2}}-V_{dx1,y1}|$
$w_{dx2,y2}(t+1) = w_{dx,y}(t) + \eta(t)*|V_{\mathrm{Node6}}-V_{dx2,y2}|$
……………………………………………………Eq. 4

Eq. 4 updates the root and the sub-nodes Node 2 and 6. η(t) is the learning rate. The adaptation of Node 2 and 6 is cascadedly propagated to the sub-structures of Node 2 and 6.

Fig. 6 illustrates the clustering result after structured clustering of patents. The attributes to train SOM are set in the top-lest area. The right map displays the clustering result and each grid contains the clustering patents shown in the left text area.



Dimension(R x W) 5 x 5
Learning Rate 20
Neighborhood 1
Iteration 10

Map Information Grid(3,3)-[275,274]

[3,3]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]

System Message

Iteration: 14
Iteration: 15
Iteration: 16
Iteration: 17
Iteration: 18
Iteration: 19
Iteration: 20
Training successful

Figure 6. SOM

4. Patent Summarization

To defense IP accusation, people need to read a lot of patents and find related ones in short time. This process costs a lot of effort and has poor productivity. Therefore, automatically summarizing patents is required to help people preview patents fast. Literature on summarization focuses on syntactic and semantic analysis to automatically produce patent summarization [10][11]. Generally speaking, the summarization process is to pick up key sentences in an original patent and eventually combine these sentences into a summarization.

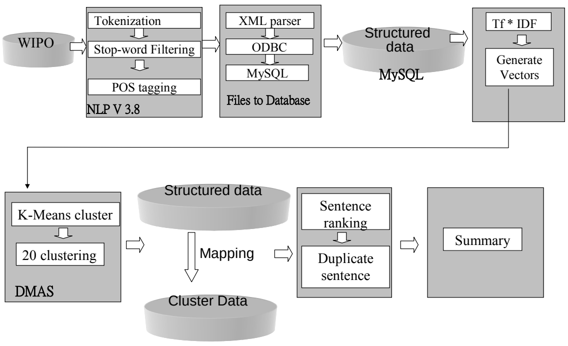


Figure 7. Patent Summarization

The proposed summarization is to extract important noun phrases according to the context near them. For example,

An influence lawmaker from the governing Labor Party on Saturday backed Spanish requests to question former Chilean dictator Gen.

The front five terms "*Labor Party on Saturday backed*" and the back five terms "*requests to question former Chilean*" determine the concept of the noun phrase "*Spanish*". Consequently, the process of summarization has five steps:

1. Find noun phrases with high score in TF*IDF.
2. Represent a noun phrase by a feature vector in front/back of N words
3. Cluster the vectors of noun phrases by k-means clustering.
4. Determine the weight of a noun phrase by comparing the clusters in step 3.
5. Sort the sentences in all documents and summarize the top-k ones with the highest weight.

Fig. 7 depicts the process of summarization. The incoming patents are processed by natural language processor and put them into a temporary database. Each feature in vector is calculated by tf*idf (as shown in Eq. 5). In step 1, stop-words are eliminated according to stop-word list. Additionally, the clustering of vectors in step 3 is to find out the concept to represent similar vectors. In step 5, the top-k sentences depend on the compression ration selected. The larger ratio produces the longer summary.

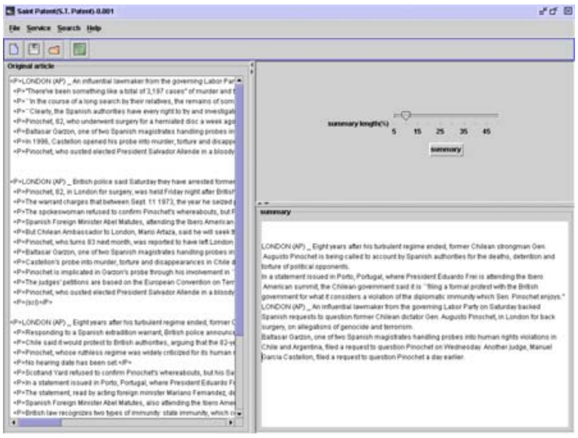
$$tf_{i,j} = \frac{freq_{i,j}}{\max freq_{l,j}} \qquad idf_i = \log_2 \frac{N}{n_i} \qquad \text{…………Eq. 5}$$

After the k-means clustering, each sentence is give a score with each cluster (as shown in Eq. 6)

$$Wp = \sum_{0}^{n} [clusters\_weight \times sentence\_weight]$$
$$= \sum_{0}^{n} [\#(cluster\_i) \times match\_term / length\_of\_sentence]$$

………………………………………………….Eq. 6

where *#(Cluster_i)* represents how many vectors in cluster I; *match_term* is the number of terms in sentence P; *length_of _sentence* indicates the length of sentence p.



<P>LONDON (AP) _ An influential lawmaker from the governing Labor Par
<P>"There've been something like a total of 3,197 cases" of murder and t
<P>``In the course of a long search by their relatives, the remains of som
<P>``Clearly, the Spanish authorities have every right to try and investigat
<P>Pinochet, 82, who underwent surgery for a herniated disc a week ag
<P>Baltasar Garzon, one of two Spanish magistrates handling probes in
<P>In 1996, Castellon opened his probe into murder, torture and disapp
<P>Pinochet, who ousted elected President Salvador Allende in a bloody

<P>LONDON (AP) _ British police said Saturday they have arrested former
<P>Pinochet, 82, in London for surgery, was held Friday night after Britis
<P>The warrant charges that between Sept. 11 1973, the year he seized
<P>The spokeswoman refused to confirm Pinochet's whereabouts, but F
<P>Spanish Foreign Minister Abel Matutes, attending the Ibero American
<P>But Chilean Ambassador to London, Mario Artaza, said he will seek t
<P>Pinochet, who turns 83 next month, was reported to have left London
<P>Baltasar Garzon, one of two Spanish magistrates handling probes in
<P>Castellon's probe into murder, torture and disappearances in Chile a
<P>Pinochet is implicated in Garzon's probe through his involvement in `
<P>The judges' petitions are based on the European Convention on Terr
<P>Pinochet, who ousted elected President Salvador Allende in a bloody
<P>(sc)</P>

<P>LONDON (AP) _ Eight years after his turbulent regime ended, former C
<P>Responding to a Spanish extradition warrant, British police announc
<P>Chile said it would protest to British authorities, arguing that the 83-ye
<P>Pinochet, whose ruthless regime was widely criticized for its human r
<P>No hearing date has been set.</P>
<P>Scotland Yard refused to confirm Pinochet's whereabouts, but his Sa
<P>In a statement issued in Porto, Portugal, where President Eduardo Fr
<P>The statement, read by acting foreign minister Mariano Fernandez, d
<P>Spanish Foreign Minister Abel Matutes, also attending the Ibero Ame
<P>British law recognizes two types of immunity: state immunity, which c

summary length(%) 5 15 25 35 45

LONDON (AP) _ Eight years after his turbulent regime ended, former Chilean strongman Gen. Augusto Pinochet is being called to account by Spanish authorities for the deaths, detention and torture of political opponents.
In a statement issued in Porto, Portugal, where President Eduardo Frei is attending the Ibero American summit, the Chilean government said it is ``filing a formal protest with the British government for what it considers a violation of the diplomatic immunity which Sen. Pinochet enjoys.''
LONDON (AP) _ An influential lawmaker from the governing Labor Party on Saturday backed Spanish requests to question former Chilean dictator Gen. Augusto Pinochet, in London for back surgery, on allegations of genocide and terrorism.
Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday. Another judge, Manuel Garcia Castellon, filed a request to question Pinochet a day earlier.

Figure 8. Patent summarization

Fig. 8 displays the result of patent summarization. The left text area shows the original patents to be summarized. Each paragraph indicates a patent. The right text area is the summarized text. The user can adjust the compression ratio to expand or shrink the summarization.

5. Patent Trend Tracking

For general users, to visualize patent knowledge using maps (called patent map) is beneficial to overview the trends of patents. Many types of patent trends can be visualized by patent maps. A majority of patent maps provide statistical information such as the statistics of patent counts with respect to companies, and the statistics of patent counts with respect to inventors [12]. However, exploiting patent maps to visualize latent knowledge like technology trends proposed in patents in some period of time is rarely studied because it is hard to obtain. To extract implicit knowledge in patents requires advanced information technologies to overcome two main tasks. The first task is to precisely extract the main techniques proposed in patents. This task relies on textual analysis to extract terms that can express the technique. The second task is to correlate the technique with time. This task relies on statistic analysis to find the significance of the technique in some period of time. Swan proposed a method to detect the significance of terms in some period of time [13]. Clustering the terms obtains the tendency of patents.

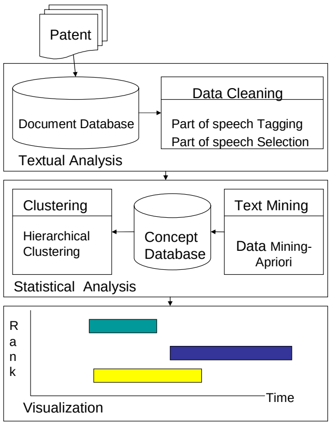


Figure 9. Patent trend tracking process

The patent trend tracking process is shown in Fig. 9. Patents are cleaned in advance and sent to "Apriori" to mine important terms [14]. In our system, the "Apriori" data mining algorithm is adapted to extract associated terms that can express techniques. Each sentence is a transaction and large item sets are mined to discover associated terms in a sentence. The associated terms are treated as concept of techniques.

The extracted terms are given significance to represent their patency. The significance is calculated by Swan'sformula in Eq. 7. [13]. Section a, b, c, d are expressed in Fig. 10.

$$x^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)} \quad \text{……………..Eq. 7}$$

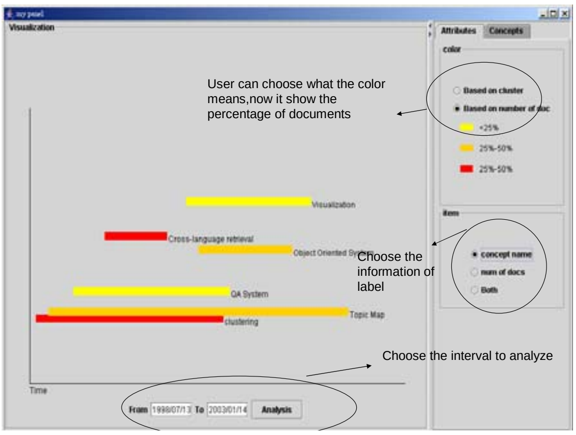
	e	ē
$t \in t_0$	a	b
$t \notin t_0$	d	d

Figure 10. The significance of terms

In Fig. 10, t represent the time and e represent the document set. The meaning of each section is as follows:

- **Section "a"** is # of documents that the term appears in time t
- **Section "b"** is # of documents that the term doesn't appear in time t
- **Section "c"** is # of documents that the term appears except time t
- **Section "d"** is # of documents that the term doesn't appear except time t

The concepts with high significance are the trend of patents in time t. As shown in Fig. 11, the interface shows patent trend with time. The X axis is the time period and the Y axis displays the significance of techniques. Different techniques are shown in different bars in each time period. The color of the technique can be set to represent the number of documents. Alternatively, the color of the techniques can also be set to the similarity among techniques. We also can set the label beside bar as name of technique concept or number of documents. Users can select the displayed techniques and time period. The patent trend tracking provides overview to patents with time and assists users to track the development of patents efficiently.


User can choose what the color means,now it show the percentage of documents

Choose the information of label

Choose the interval to analyze

○ Based on cluster
● Based on number of doc

● concept name
○ num of docs
○ Both

From 1998/07/13 To 2003/01/14 Analysis

Figure 11. Paten Trend Tracking

6. Conclusions

In this paper, a patent retrieval system is developed to annotate and discover knowledge from patents. The system contains an on-line retrieval module, structured patent clustering, patent summarization and patent trend tracking. In the on-line retrieval module, we apply ontology-based search and use concepts to extend search scope. We propose a structure model to analyze retrieved patent and apply structured clustering to patents. The clustering result is summarized according to the context of important noun phrases. Moreover, a visualization interface of for tracking patent trends is designed to overview the trends of patents developed in a period of time.

The effectiveness of the proposed system relies on precise textual analysis. In both retrieval and trend tracking, the extraction of concept depends on a statistical algorithm and has limitation in the performance. We plan to adapt natural language processing like POS tagging, corpus training to improve the accuracy of concept extraction. Moreover, SOM clustering with structure of patents consumes more time and map space than traditional SOM. The future work is to reduce the map by adjusting the training weight of adaptation.

7. Acknowledgement

This research was supported by the Software Technology for Advanced Network Application project of Institute for Information Industry and sponsored by MOEA , R.O.C.

References

[1] *ACL Workshop on Patent Corpus Processing*, Sappora, Japan, 2003. Available at http://acl.ldc.upenn.edu/acl2003/patent/index.htm.

[2] *ACM SIGIR Workshop on Patent Retrieval*, Athens, Greece, 2000. Available at http://research.nii.ac.jp /ntcir/sigir2000ws/.,

[3] L.S. Larkey, "A patent search and classification system," *Proceedings of the fourth ACM conference on Digital libraries*, , pp.179-187, 1999.

[4] A. Fujii and T. Ishikawa, "Document Structure Analysis in Associative Patent Retrieval", *NTCIR Workshop 4 Meeting Working Notes*. Available at http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/

[5] A. Shinmori, M. Okumura, Y. Marukawa and M. Iwayama, "Can Claim Analysis Contribute toward Patent Map Generation", *NTCIR Workshop 4 Meeting Working Notes*. Available at http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/

[6] B.Hammer, B.J.Jain, "Neural methods for non-standard data", , *European Symposium at Artificial Neural Networks'2004*, D-side publications, Verleysen, pp. 281-292, 2004.

[7] Alessandro Sperduti, "Neural Networks for Adaptive Processing of Structured Data", *Lecture Notes in Computer Science*, pp. 5-12, 2001.

[8] Alessandro Sperduti and Antonina Starita, "Supervised Neural Networks for Classification of Structures", *IEEE transaction on Neural Networks*, 8(3), pp. 714-735, 1997.

[9] Markus Hagenbuchner, Alessandro Sperduti and Ah Chung Tsoi, „A Self-Organizing Map for Adaptive *Processing of Structured Data", IEEE Transactions on Neural Networks*, 14(3), pp. 491-505, 2003.

[10] E. Hovy, and C. Y. Lin (1999), "Automated Text Summarization in SUMMARIST," In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization*, MIT Press, pp. 81-94, 1999.

[11] J. Y. Yeh, H. R. Ke, and W. P. Yang (2002), "Chinese Text Summarization Using A Trainable Summarizer and Latent Semantic Analysis," *Proceedings of the 5th International Conference on Asian Digital Libraries*, Singapore, 2002

[12] PatentGuider, Available at http://www.learningtech .com.tw/products/pg_function .aspx#

[13] R.Swan and J.Allan, "Automatic generation of overview timeliness," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athen, Greece, 2000

[14] J. Han and M. Kamber, *Data mining: concepts and techniques*, San Mateo, CA, Morgan Kaufmann Publishers, 2001.