TEXT MINING FOR PATENT MAP ANALYSIS

Yuen-Hsien Tseng, Fu Jen Catholic University, Taiwan, tseng@lins.fju.edu.tw
Yeong-Ming Wang, LungHwa Uni. of Sci. and Tech., Taiwan, yeongmin@ms37.hinet.net
Dai-Wei Juang, WebGenie Information LTD., Taiwan, david@webgenie.com.tw
Chi-Jen Lin, WebGenie Information LTD., Taiwan, dan@webgenie.com.tw

ABSTRACT

Patent documents contain important research results. However, they are lengthy and rich in technical and legal terminology such that it takes a lot of human efforts to analyze them. Automatic tools for assisting patent analysis are in great demand. This paper describes some methods for patent map analysis based on text mining techniques. We experiments on a real-world patent map created for an important technology domain: "carbon nanotube". We show that most important category-specific terms occur in the machine-derived extracts from each patent segment and that machine-derived feature terms can be as good as those selected manually. The implications is that future patent mapping can be done in a more effective and efficient way.

Keywords: Patent map, text mining, carbon nanotube, feature extraction, patent analysis

INTRODUCTION

Patent documents contain important research results. If carefully analyzed, they can show technology details and relations, reveal business trends, or inspire novel industrial solutions. In recent years, patent analysis had been recognized as an important task at the government level. Public institutions in Taiwan, Korean, and Japan have invested resources in the training and the performing of the task [1-2]. However, patent documents are often lengthy and rich in technical and legal terminology and are thus hard to read and analyze for non-specialists. Automated technologies for assisting patent analysis are in great demand.

A patent document contains dozens of items for analysis; some are structured, meaning they are uniform in semantics and in format across patents such as patent number, filing date, or inventors; some are unstructured, meaning they are free texts such that they are quite different in length and content for each patent such as claims, abstracts, or descriptions of the invention. The visualized results of patent analysis are, in this paper, called *patent graphs* if they are from the structured data and *patent maps* if they are from the unstructured texts, although, loosely speaking, patent maps can refer to both cases [3]. For example, Table 1 shows part of a patent map created by analyzing 95 patent documents [4]. These patents, issued before February 19, 2002, are the search results of the keyword: "carbon nanotube" from the database of USPTO [5]. In Table 1, manual-assigned categories regarding the *technologies* of the patents are listed in rows and those regarding the *effects* of the patents are listed in columns. The patents are then assigned to the cells of the technology-effect matrix based on their textual contents.

Creating and updating such a map requires a lot of human efforts. As in the above "carbon nanotube" map (called *CNT map*), five specialists spent more than one months in analyzing about one hundred patents. Although it may go unnoticed, such efforts indeed involve some text

mining processes such as text segmentation, summary extraction, keyword identification, topic detection, taxonomy generation, term clustering, and document categorization. Previous studies have shown potential usefulness of these techniques in assisting the task of patent analysis and patent mapping [6-11].

Table 1. Part of the technology-effect matrix for "Carbon Nanotube" from 95 US patents.

Technology \ Effect		Material	Performance		Product
		Carbon nanotube	Purity	Electricity	FED
Manufacture	Gas reaction	5346683 6129901 …	6181055 6190634	6221489	6232706
	Catalyst	5424054 5780101 …	6333016		6339281
	Arc discharging	5424054	6190634 6331262	5916642	5916642
Application	Display			6346775	5889372 5967873 …

In this paper, we apply some of the text mining techniques to certain sub-tasks that have a clear goal. Specifically, we would like to know where the important category-specific terms locate in a patent document. Knowing this not only allows analysts to spot the relevant segments more quickly for classifying patents in the map, but also provides insights to possibly improve automated clustering and/or categorization in creating the map.

In the next section, we introduce the method for selecting important terms for classification. We then describe how we analyze where the important term often occur in a patent by using the classified patents in the above patent map as our experimental data. Specifically, we divide each patent document into several textual segments according to their content. The segments where important terms occur are recorded and such occurrences are accumulated. The location distribution of important terms selected automatically is then compared to the distribution of those selected manually. We also experiment on how automated summarization helps in identifying important category-specific terms. Finally, we conclude this paper with our main findings, implications, and possible future work.

FEATURE SELECTION

Selecting the best category-specific terms, or called *category features* hereafter, for document categorization have been studies in the fields of machine learning and information retrieval. Yang et al [12] had compared five different methods. They found that Chi-square is among the best that lead to best performance in the task of text categorization. The Chi-square method computes the relatedness of term T with respect to category C according to the $x^2()$ formula in Table 2, where TP (True Positive) , FP (False Positive), FN (False Negative), and TN (True Negative) denote the number of documents that belong or not belong to C and that contain or not contain T, as shown in Table 3. However, Chi-square weights positive and negative related terms equally. A remedy to this is the correlation coefficient *co()*, which is also called one-sided chi-square, as shown in Table 2.

Ng et al [13] pointed out that correlation coefficient selects exactly those words that are highly indicative of membership in a category, whereas the chi-square method will not only pick out this set of terms but also those terms that are indicative of nonmembership in that category. This is especially true when the selected terms are in small number. As an example, in a small real-world collection of 116 documents with only two exclusive categories: construction vs. non-construction, some of the best and worst terms that are computed by chi-square and correlation coefficients are shown in Table 4. As can be seen, due to the square nature, the chi-square weights negative related terms as high as positive related terms. (For the term: 'engineering', the square of -0.7880 is 0.6210.)

Table 2. Chi-square and correlation coefficient

$$\chi^2(T,C) = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})^2}{(\mathrm{TP} + \mathrm{FN})(\mathrm{FP} + \mathrm{TN})(\mathrm{TP} + \mathrm{FP})(\mathrm{FN} + \mathrm{TN})}$$

$$Co(T,C) = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{FP} + \mathrm{TN})(\mathrm{TP} + \mathrm{FP})(\mathrm{FN} + \mathrm{TN})}}$$

Table 3. Term-Category distribution

		Term T	
Category C		Yes	No
	Yes	TP	FN
	No	FP	TN

Table 4. Some best and worst terms computed by Chi-square and correlation coefficient in a collection with two exclusive categories.

Chi-square				Correlation coefficient			
Construction		Non-Construction		Construction		Non-Construction	
engineering	0.6210	engineering	0.6210	engineering	0.7880	equipment	0.2854
improvement	0.1004	improvement	0.1004	improvement	0.3169	procurement	0.2231
…				…			
kitchen	0.0009	kitchen	0.0009	communiqué	-0.2062	improvement	-0.3169
update	0.0006	update	0.0006	equipment	-0.2854	engineering	-0.7880

DOCUMENT SETS

The textual content of a patent document from USPTO contains title, abstract, claims, and description. The description, the main body of the detailed content, often have sub-sections with titles in uppercase and in a single line, such as FIELD OF THE INVENTION, BACKGROUND, SUMMARY OF THE INVENTION, and DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT, although some patents may have more or less such sub-sections or have slightly different title names. As the titles show, each section or sub-section plays a different role in describing a patent. When analyzing the content for patent mapping, each section was utilized differently for different category assignment. Especially, the section that a term occurs may determine how important the term could be and how it could be used.

Using regular expression matching, all these sections can be segmented apart with a high success rate [14]. From the 92 patents in the CNT map, we segmented each patent and created a document set for each segment such that each set contains 92 documents with each document having the corresponding segment of a patent as its content. Since patent description can be very lengthy, the excessive details may distract analysts from topic identification. Thus except these segment sets, we also created two other sets for comparison. One is the segment extraction set;

the other is the full set. All these document sets are named for convenience and defined as follows:

abs: corresponds to the 'Abstract' section of each patent
app: corresponds to the segment of FIELD OF THE INVENTION
task: corresponds to the BACKGROUND OF THE INVENTION
sum: corresponds to the SUMMARY OF THE INVENTION
fea: DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT
cla: corresponds to the claims section of each patent
seg_ext: corresponds to the top-ranked sentences of each document from the sets: abs, app, task, sum, and fea.
full: corresponds to all the documents from the sets: abs, app, task, sum, and fea.

The extracted sentences in the seg_ext set are those ranked based on traditional automated summarization techniques [14]. For examples, sentences which contain more keywords or title words are weighted more. Sentences which occur at some locations such as the starting and ending paragraphs or the starting and ending sentences of some paragraphs are weighted heavier. The sentences are then ranked by their total weights and the best six sentences from individual segments of a patent are gathered to constitute a document in the seg_ext set.

Our regular expression matching rule is quite effective such that there are only one empty document in the 'app' set, two in the 'task' set, five in the 'sum' set, and one in the 'fea' set. Among these nine empty documents, five are due to the lack of such sections in the original patent documents, three are due to the use of special section titles, and one is due to the erroneous spelling of the section title.

EXPERIMENTS AND RESULTS

In the technology-effect matrix of the CNT map, there are 9 leaf-categories in the technology taxonomy and 21 leaf-categories in the effect taxonomy. All these leaf-categories are grouped together based on more general topics as shown in Table 1. We analyze the important category features (*ICFs*) based on the correlation coefficient described above. *N* (*N*=50) top-ranked features for each of the leaf-category in each document set were extracted. The number of sets that such a feature occurs, denoted as *sc* for *set count*, and the sum of correlation coefficients accumulated over all such sets, denoted as *ss* for *set sum*, were calculated. The features were then ranked by *sc* in descending order. An example for some ranked features of the category FED (Field Emission Display) is shown in Table 5. The second row shows that the feature: 'emit' occurs in the entire 8 document sets. It occurs in 12 documents in the abs set and has a correlation coefficient of 0.62 in that set. The first column titled *rel* denotes whether the term is judged as a relevant category feature by one of the authors who participated in the creation of the CNT map. It should be noted that such judgment is quite subjective and sometimes contradictive such that good features that are effective in discriminating the categories of current analyzed documents may not be judged as relevant. This is often observed in feature selection studies as statistically important terms are hard to identify manually by the meaning of the term only.

The *M* best features ranked by segment count were then further aggregated in each document set for each category. Part of the results for *M*=30 are shown in Table 6. As the third row shows (the

FED category), among the 30 best features, 16 occur in the abs set, covering 53% of them, and 21 in the seg_ext set, covering 70.0% of the 30 terms, which reveals that most ICFs can be found in the best six sentences of each segment. The 30 best features ranked by segment count were further checked by one of the CNT map creators. The aggregated results are shown in Table 7. The third row shows that among 7 relevant features, 6 of them occur in the 'abs' set, covering 87.5% of them.

Table 5. Some feature terms and their distribution in each set for the category FED.

rel	term	sc	ss	abs		app		Task		sum		fea		cla		seg_ext		full	
	emit	8	4.86	12	0.62	13	0.58	21	0.55	17	0.59	22	0.70	14	0.61	20	0.63	27	0.59
yes	emission	8	5.07	20	0.69	17	0.59	31	0.62	21	0.73	34	0.63	20	0.63	33	0.64	40	0.54
yes	display	8	5.06	9	0.50	12	0.62	22	0.64	14	0.61	24	0.71	10	0.62	23	0.68	34	0.68
	cathode	8	3.86	12	0.39	9	0.42	27	0.48	14	0.54	30	0.53	15	0.51	25	0.52	41	0.47
	pixel	7	3.14	3	0.33			8	0.46	3	0.33	12	0.62	2	0.27	5	0.43	17	0.72
	screen	5	1.74	2	0.27	2	0.27	8	0.37			18	0.43					19	0.41
yes	electron	5	1.71	27	0.31	25	0.40			36	0.28			27	0.37	61	0.35		
yes	voltage	4	1.48					20	0.45			45	0.37			16	0.28	52	0.39

Table 6. Occurrence distribution of 30 top-ranked terms in each set for some categories.

category	T_No	abs	App	Task	sum	fea	cla	seg_ext	full
Carbon nanotube	30	15/50.0%	12/40.0%	14/46.7%	20/66.7%	13/43.3%	19/63.3%	20/66.7%	13/43.3%
FED	30	16/53.3%	14/46.7%	22/73.3%	19/63.3%	21/70.0%	19/63.3%	21/70.0%	22/73.3%
device	30	21/70.0%	17/56.7%	9/30.0%	16/53.3%	7/23.3%	19/63.3%	17/56.7%	8/26.7%
Derivation	30	14/46.7%	6/20.0%	7/23.3%	11/36.7%	8/26.7%	13/43.3%	13/43.3%	11/36.7%
electricity	30	12/40.0%	10/33.3%	10/33.3%	10/33.3%	8/26.7%	8/26.7%	13/43.3%	12/40.0%
purity	30	12/40.0%	12/40.0%	7/23.3%	20/66.7%	9/30.0%	17/56.7%	18/60.0%	14/46.7%
High surface area	30	19/63.3%	13/43.3%	13/43.3%	17/56.7%	8/26.7%	9/30.0%	16/53.3%	8/26.7%
magnetic	30	18/60.0%	11/36.7%	6/20.0%	14/46.7%	14/46.7%	13/43.3%	15/50.0%	13/43.3%
energy storage	30	16/53.3%	17/56.7%	13/43.3%	17/56.7%	6/20.0%	10/33.3%	21/70.0%	12/40.0%

Table 7. Occurrence distribution of manually ranked terms in each set for some categories.

category	T_No	abs	app	task	sum	fea	cla	seg_ext	full
Carbon nanotube	4	3/75.0%	2/50.0%	2/50.0%	3/75.0%	2/50.0%	2/50.0%	3/75.0%	2/50.0%
FED	7	6/85.7%	6/85.7%	6/85.7%	4/57.1%	6/85.7%	4/57.1%	6/85.7%	5/71.4%
device	2	2/100.0%	1/50.0%	0/0.0%	1/50.0%	1/50.0%	2/100.0%	1/50.0%	0/0.0%
electricity	2	2/100.0%	2/100.0%	2/100.0%	2/100.0%	1/50.0%	2/100.0%	0/0.0%	0/0.0%
purity	2	2/100.0%	2/100.0%	0/0.0%	1/50.0%	1/50.0%	1/50.0%	0/0.0%	1/50.0%
High surface area	8	6/75.0%	2/25.0%	3/37.5%	5/62.5%	1/12.5%	2/25.0%	4/50.0%	1/12.5%
magnetic	5	3/60.0%	1/20.0%	2/40.0%	1/20.0%	3/60.0%	0/0.0%	4/80.0%	3/60.0%
energy storage	2	2/100.0%	2/100.0%	1/50.0%	2/100.0%	1/50.0%	1/50.0%	2/100.0%	0/0.0%

Table 8. Occurrence distribution of terms in each segment averaged over all categories.

Taxonomy	Set		abs	app	task	sum	fea	Cla	seg_ext	full
	nc	nt								
Effect	9	M=30	52.96%	41.48%	37.41%	53.33%	34.81%	47.04%	**57.04%**	41.85%
Effect*	8	4	**86.96%**	66.34%	45.40%	64.33%	51.03%	54.02%	55.09%	30.49%
Tech	21	M=30	49.37%	25.56%	26.51%	**56.51%**	34.44%	46.51%	**56.03%**	40.95%
Tech*	17	4.5	**59.28%**	29.77%	23.66%	49.43%	34.46%	**60.87%**	44.64%	32.17%

Table 9. Maximum correlation coefficients in each set averaged over all categories.

Taxonomy	Set		abs	app	task	sum	fea	cla	seg_ext	full
	nc	nt								
Effect	9	M=30	**0.58**	**0.49**	**0.54**	**0.55**	**0.55**	**0.57**	**0.56**	**0.55**
Effect*	8	4.0	0.52	0.43	0.39	0.52	0.48	0.47	0.44	0.33

Tech	21	M=30	**0.64**	**0.58**	**0.65**	**0.66**	**0.66**	**0.67**	**0.68**	**0.68**
Tech*	17	4.5	0.47	0.29	0.34	0.44	0.35	0.51	0.43	0.42

The term-covering percentages (of the best terms and the relevant terms) of each set were accumulated and averaged over all categories with respect to the technology taxonomy and the effect taxonomy, respectively. The results are shown in Table 8. The second column denotes the number of categories (*nc*) in that taxonomy and the third column denotes the average number of terms in each category (*nt*) for calculating the term-covering average. The rows with a star in the first column denote that the average is calculated from the human-judged relevant terms. As the bold-faced data show, most automated-derived ICFs occur in the segment extracts, while most human-judged ICFs occur in the abstract section of a patent document.

In Table 9, we show the averages of the maximum correlation coefficients of the 30 best terms and of the relevant terms calculated over all categories in each taxonomy for each document set. Note: the higher the coefficients, the better the term in predicting a document's categories. As shown in Table 9, the coefficients of automated-derived ICFs are all higher than those of human-judged ICFs.

To see if important sets change when the number of best terms (ranked by segment count) changes, we varied the number M for additional values, 10 and 50, and calculated the averaged term-covering rates for each set. The results in Figure 1 show that 'abs', 'sum' and 'seg_ext' are important sets. The 'full' set becomes important only when more terms are included for counting.

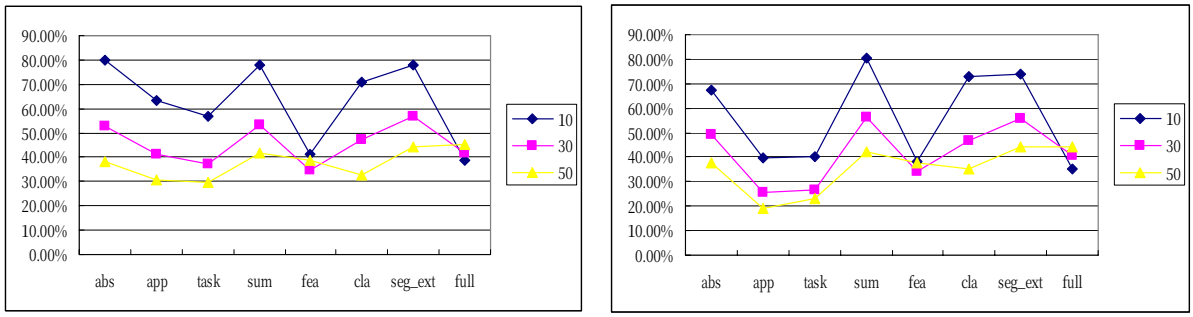

Figure 1. Term-covering rates for M best terms, where M=10, 30, and 50. Left figure is for the effect taxonomy and right figure is for the technology taxonomy.

From the above results, we summarize our findings as follows: (1) Most ICFs ranked by correlation coefficient occur in the segment extracts, the Abstract section, and the SUMMARY OF THE INVENTION section. (2) Most ICFs selected by humans occur in the Abstract section or the Claims section. (3) The segment extracts lead to more top-ranked ICFs than the full texts, regardless whether the category features are selected manually or automatically. (4) The ICFs selected automatically have higher capability in discriminating a document's categories than those selected manually according to the correlation coefficient.

CONCLUSION

Our findings confirm some of our speculations before we conduct this experiment. First, text summarization techniques help in patent analysis and organization, either automatically or manually. This is because patents tend to be lengthy, rich in terminology, and full of details, which may distract the topic analysis for humans and for machines alike. Second, machine-derived ICFs may be as good as or even better than those derived manually. This can be seen by re-inspecting individual terms and by the higher coefficients they have as shown in Table 9. Third, patent abstracts are shown to be very important in Table 8 and 9 in patent analysis. There was once a rumor in the field of patent analysis saying that patent abstracts were not reviewed by patent officers such that the contents of the abstracts were not reliable.

Our findings indeed need further verifications. For example, we can use these ICFs for classifying the patents to see if they are as discriminative as the correlation coefficient predicts. Also the segment extracts should be compared with the full texts and other segments in the real task of classification to show its advantages. Our future work will conduct more experiments to verify these findings.

Since the patent map has been created in 2002, the number of patents about “carbon nanotube” has increased from 95 to more than 400. The map needs to be updated to reflect the latest technology trend. With the lessons and knowledge learned from the experiments in this paper, we expect that the map can be updated in a more efficient way to help decide the research strategies on the fast-growing “carbon nanotube” industry.

REFERENCES

1 Billy Chen, ”Introduction to Patent Map”, Lecture Notes for the training of patent mapping and patent analysis, Taipei, National Science Council, 1999. (in Chinese)
2 Young-Moon Bay, “Development and Applications of Patent Map in Korean High-Tech Industry” The first Asia-Pacific Conference on Patent Maps, Taipei, Oct. 29, 2003, pp. 3-23.
3 Shang-Jyh Liu, “Patent Map - A Route to a Strategic Intelligence of Industrial Competitive-ness,” The first Asia-Pacific Conference on Patent Maps, Taipei, Oct. 29, 2003, pp. 2-13.
4 Fu-Der Mai, Fengtai Hwang, Kuo-ming Chien, Yeong-Ming Wang, Chiu-yen Chen, *Patent map and analysis of Carbon nanotube*, Science and Technology Information Center, National Science Council, R.O.C., April 2002.
5 “United States Patent and Trademark Office”, http://www.uspto.gov/
6 M. Iwayama, A. Fujii, N. Kando, and A. Takano, " Overview of Patent Retrieval Task at NTCIR-3," Proceedings of the 3rd NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan.
7 NTCIR-4 Patent Retrieval Task, http://www.slis.tsukuba.ac.jp/~fujii/ntcir4 /cfp-en.html.
8 Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa and Makoto Iwayama, “Patent Claim Processing for Readability - Structure Analysis and Term Explanation,” Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan.
9 Svetlana Sheremetyeva, “Natural Language Analysis of Patent Claims,” Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan.

10 "Intelligent Patent Analysis through the Use of a Neural Network: Experiment of Multi-Viewpoint Analysis with the MultiSOM Model," Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan, pp. 7-23.
11 C. J. Fall, A. Torcsvari, K. Benzineb, G. Karetka, "Automated Categorization in the International Patent Classification," ACM SIGIR Forum, Vol. 37, No. 1, 2003, pp. 10 - 25.
12 Y. Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. of the Intern. Conference on Machine Learning, 1997, pp. 412-420.
13 Hwee Tou Ng, Wei Boon Goh and Kok Leong Low, "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization," Proc. of the 20th International Conference on Research and Development in Information Retrieval, 1997, Pages 67 - 73.
14 Yuen-Hsien Tseng, "Knowledge Discovery in Patent Texts: Techniques and Challenges," Conf. on Modern Information Organization and Retrieval, Taipei, Nov. 19, 2004, pp. 111-124. (in Chinese)