Patent Retrieval System Using Document Filtering Techniques

Naomi INOUE Kazunori MATSUMOTO Keiichiro HOASHI Kazuo HASHIMOTO
KDD R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka, Saitama 356-8502 JAPAN
+81-492-78-7397
{inoue,matsu,hoashi,kh}@kddlabs.co.jp

Abstract

Existing patent retrieval systems are difficult to use by amateur users because of several drawbacks. We have developed a novel type of patent retrieval system which anybody can make use of easily, and have been using it in our laboratories since the beginning of this year. Our system uses document filtering techniques. These filtering techniques are based on a probabilistic model which searches documents relevant to the user's interest. This paper describes our patent retrieval system, filtering method and experimental results.

1 INTRODUCTION

Traditional patent retrieval is considered to be a task which documents relevant to a user's interest are retrieved from large stores of textual data. Existing patent retrieval systems require the user to input IPC (International Patent Classification) codes in order to select relevant document sets. The systems then retrieve documents containing the user's search keywords from the selected document sets. However, this retrieval process has the following drawbacks.

(1) The user has to be familiar with IPC codes.

(2) Much time is necessary until the user obtains relevant documents because he/she has to browse an overwhelming amount of non-relevant documents.

(3) A user who wants to monitor documents relevant to specific information must repetitively input search keywords to reject useless documents. Moreover, the user can not discriminate between new and past-retrieved documents.

These drawbacks make it difficult for an amateur user to use traditional patent retrieval systems to obtain relevant patents.

We have developed a novel type of patent retrieval system which anybody can make use of easily. Document filtering techniques are used in our system. Document filtering is a task which monitors the flow of incoming documents and selects those which the system regards as relevant to the user's interest. The user's interest is expressed within the system as a profile. Our system calculates the similarity between the profile and each incoming document, and then send documents with higher similarity than a certain threshold to the user by e-mail. Therefore, IPC codes are not used in the retrieval process. This means the user does not have to be familiar with IPC codes. We believe this system has a high level of usability.

Section 2 of this paper describes the structure of our patent retrieval system, section 3 describes the filtering method, section 4 gives experimental results and section 5 presents our conclusions .

2 System Structure

The structure of our patent retrieval system is shown in Figure 1. We receive patent data from a patent office every 2 weeks. The amount of incoming documents are 3,000 documents(400MB) on average. The patent database of our patent retrieval system is updated after the receiving terminal finishes receiving documents. Our system then calculates the similarity between a user's profile and each incoming document, and send documents with higher similarity than a certain threshold to the user. Currently, 157 users are using our patent retrieval system.

All the user has to do is to register his/her own profile by using an Internet browser such as Netscape or Internet Explorer. The profile consists of text files such as patents which the user already applied or documents in which the user's interest is described. The method of registering the user's profile is very simple. The user

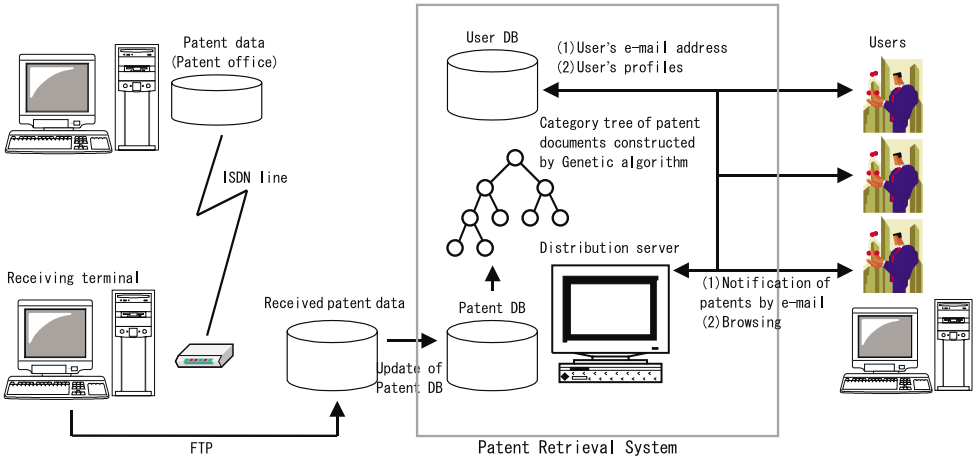

Figure 1: Structure of our Patent Retrieval System

selects only interesting patents from a list of patents already applied by the user. The user can also register his/her interest by writing text in a specific column of a registration form and then submitting the form.

3 Filtering Method

In many document filtering systems, the user's profile and incoming documents are indexed by a vector space model. The similarity between the user's profile and each incoming document is then usually calculated as an inner-product or cosine of two vectors.

On the other hand, probabilistic models have been exploited in information retrieval and text categorization because these models have a solid formal grounding in probabilistic theory. The adaptation of probabilistic models to document filtering is straightforward. To retrieve documents relevant to a user's interest, a simple strategy is to calculate the similarity between each incoming documents and all user profiles. This exhaustive search increases computational time in proportion to the product of the number of user's profiles stored in the system and the number of incoming documents. In general, probabilistic models have high accuracy but need a large amount of computational time for the calculation of the similarity. Therefore, there would be a limit to the exhaustive search as the number of user's profiles stored in the system or the number of incoming documents increases. That is to say, an exhaustive search based on probabilistic models can be used in a filtering system of reasonable scale, but is considered to be unsuitable for a large scale filtering system. However, Iwayama reported that his proposed cluster-based search with probabilistic clustering algorithm was effective, efficient and noise tolerant for retrieval from a large amount of data[1]. For this reason, we designed our patent retrieval system to perform both an exhaustive search and a cluster-based search. The exhaustive search is performed if both the number of user's profiles stored in the system and the number of incoming documents are reasonable, and the cluster-based search is performed if the number of user's profiles stored in the system and/or the number of incoming documents is large.

In this section, we first describe the method of exhaustive search based on a probabilistic model, and then describe the method of cluster-based search.

3.1 Calculation of Similarity between Two Documents

A document filtering system based on probabilistic models calculates a posterior probability $P(c|d)$, the probability that a user's profile d is classified into a cluster c.

Many methods of calculating posterior probability have been proposed[2][3][4][5]. Our patent retrieval system adopts Iwayama's formulation because it has the following advantages over other calculation methods.

(1) it considers within-document term frequencies.

(2) it considers term weighting for incoming documents.

(3) it is less affected by having an insufficient number of training documents.

Iwayama's formulation is described as follows.

$$P(c|d) = P(c)\sum_{t}\frac{P(T=t|c)P(T=t|d)}{P(T=t)} \qquad (1)$$

"T=t" means that a randomly selected term T from the document d is equal to t. Probabilities on the right-hand side of this equation are estimated as follows:

- $P(T = t|d)$: relative frequency of a term t in a user's profile document d.
- $P(T = t|c)$: relative frequency of a term t in a cluster c.
- $P(T = t)$: relative frequency of a term t in the entire set of incoming documents.
- $P(c)$: relative frequency of documents that belong to c in the entire set of incoming documents.

The similarity between a user's profile and each incoming document is calculated by the following equation. In this equation, d_x corresponds to a user's profile and d_y corresponds to an incoming document.

$$Sim(d_x, d_y) = \frac{P(\{d_x, d_y\}|d_x) \cdot P(\{d_x, d_y\}|d_y)}{P(\{d_x\}|d_x) \cdot P(\{d_y\}|d_y)} \quad (2)$$

The right-hand side of this equation corresponds to the ratio of the posterior probability after d_x and d_y are merged into a cluster and the posterior probability before they are merged. This equation can be derived easily from equation(5).

3.2 Cluster-Based Search

Iwayama proposed a hierarchical clustering algorithm that constructs a set of clusters having maximum Bayesian posterior probability[6]. This algorithm is called Hierarchical Bayesian Clustering(HBC).

When the number of documents is N, the calculation amount required to make a cluster is O(N^2), therefore it is extremely difficult to make a cluster of a large number of documents by conventional systems.

In this subsection, we give the HBC algorithm first, and then describe our proposed clustering techniques[7] in order to improve the clustering speed.

3.2.1 HBC Clustering Algorithm

Initially, each document belongs to a cluster whose only member is the document itself. For every pair of clusters, HBC calculates the increase of posterior probability after the pairs are merged, selects the pair that results in the maximum increase, and those clusters are then merged to form a new cluster.

To see the details of this merge process, consider a merge step k+1($0 \leq k \leq N-1$). In the step k+1, a data collection of N, $D = \{d_1, d_2, ..., d_N\}$, has been partitioned into a set of clusters $C_k = \{c_1, c_2, ...\}$. That is, each datum $d_i \in D$ belongs to a cluster $c_j \in C_k$. The overall posterior probability at this point becomes

$$P(C_k|D) = \prod_{c_j \in C_k} \prod_{d_i \in c_j} P(c_j|d_i) \quad (3)$$

The set of clusters C_k is updated as follows:

$$C_{k+1} = C_k - \{c_x, c_y\} + \{c_x \cup c_y\} \quad (4)$$

After the merge, the posterior probability is updated as follows:

$$P(C_{k+1}|D) = P(C_k|D) \frac{\prod_{d_i \in c_x \cup c_y} P(c_x \cup c_y|d_i)}{\prod_{d_i \in c_x} P(c_x|d_i) \prod_{d_i \in c_y} P(c_y|d_i)} \quad (5)$$

When clustering is performed with the above algorithm, the number of calculations of evaluation values of the posterior probability is:

$$_N C_2 + \sum_{k=1}^{N-2} k = (N-1)^2 = O(N^2) \quad (6)$$

where N is the number of documents. Thus, the calculation amount is $O(N^2)$. Hence when a large number of documents is handled, cluster generation becomes difficult.

3.2.2 Proposed Clustering Method

Iwayama proposed an approximate clustering technique for applying HBC to a large number of document sets[8]. The basic idea of the approximation is to decrease processing time in deciding a layer by computing the similarity from selected documents instead of all documents. However, in Iwayama's proposed method, the layer is not always optimum because documents are selected at random.

We proposed a new approximate clustering algorithm that improved the precision of Iwayama's method. We proposed selecting documents by applying a genetic algorithm (referred to hereafter as GA)[9] for deciding a quasi-optimum layer and using a MDL criteria for evaluating the layer structure of a cluster tree. Our method gives better accuracy than Iwayama's method, because the layer structure of a cluster tree constructed by our method is quasi-optimum. The advantage of the GA-based algorithm is that it is known to converge speed compared with other optimal methods.

To improve the precision of Iwayama's clustering method, we propose the following method, combining conventional strict clustering with top-down clustering using GA. Assume that the total number of documents to be clustered is N, and the number of documents within a range which can be handled by a strict clustering method is M.

```
procedure GA-clustering()
    all documents are assigned to a root document set (D_root);
    D_root is registered as cue-Q;
    while (Q is not empty) {
        a document set D_p at the head of Q is extracted;
        if (the number of documents |D_p| < M)
            HBC(D_p); /* clustering of D_p by HBC */
        else {
            D_d = Select(D_p);
            /* document set D_d, constructed from M documents
            considered to be optimum, which are extracted
            from D_p.
            The coding length of a cluster is minimized based on
            an MDL criteria.
            A genetic algorithm is used for the analytical
            search. */
            HBC(D_d);
            The remaining documents (D_p − D_d) are
            assigned to the nearest leaf (L_i ∈ C_d);
            Document sets assigned to the D_i = L_i;
            if(number of documents |D_i| > 0)
                D_i is added to Q;
        }
    }
endproc
```

3.2.3 Evaluation of Proposed Clustering Method

Aoki reported that the proposed clustering method reduces the number of merges from $O(N^2)$ to $O(N)$, and the time required for one merge is substantially constant regardless of N. Moreover, it was found that the precision of the proposed method is higher than that of Iwayama's approximation clustering method.

Detailed experimental results are described in Aoki[7].

3.2.4 Cluster-Based Search Method

If the number of user's profiles stored in the system and/or the number of incoming documents is large, exhaustive search can not be used because of its computational time. In this case, cluster-based search should be used.

In this search, a cluster tree is constructed from a large amount of patent documents including user's profiles. Each incoming document is compared with each cluster from top of the tree and is assigned to the nearest leaf of the cluster tree. Then the K-nearest neighbor documents to each incoming document are retrieved.

4 Experiments

As mentioned in section2, the number of users of our system is 157 and about 3,000 documents come in every 2 weeks. These numbers are considered to be reasonable

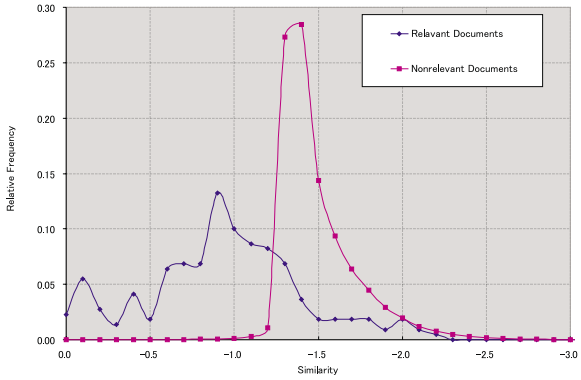

Figure 2: Experimental Result

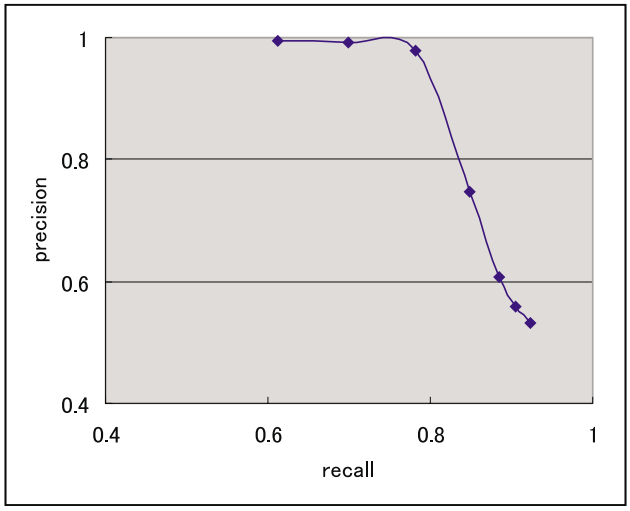

Figure 3: Recall-Precision

for exhaustive search. In this section, we describe the experimental results for this exhaustive search.

We used 21 patents as user's profiles, and had a professional organization conduct a search for similar patents in 10,000 patents extracted at random from patents published between 1993 and 1997. Experimental results are shown in Figure 2.

The result shows that most documents with high similarity over -1.2, are relevant and relevant documents can be distinguished from nonrelevant documents.

A recall and precision curve is shown in Figure 3. We define precision as the ratio of documents that are truly relevant to those that are classified as relevant, and recall as the ratio of truly relevant documents that are classified as relevant.

Figure 3 shows that precision decreases significantly as the recall increases.

5 Conclusion

In this paper, we described our patent retrieval system. Our system has the following characteristics.

(1) **Good Usability:** An IPC code is not used in the retrieval process. This means the user does not have to be familiar with IPC codes.

(2) **Easy Accessibility:** all the user has to do is to register his/her own profile by using an Internet browser such as Netscape or Internet Explorer. The profile consists of text files such as patents which the user has already applied and documents in which the user's interest is described.

(3) **High Retrieval Accuracy:** Document filtering techniques are based on a probabilistic model and offer high accuracy. Moreover, a cluster-based search using a fast document categorization method is provided.

We are planning to improve the accuracy by using the relevance feedback and calculating the similarity between a user's profile and each incoming document by combining several criterion.

References

[1] M. Iwayama, T. Tokunaga: *"Cluster-Based Text Categorizarion: A Comparison of Category Search Strategies"*, Proceedings of SIGIR-95, pp.273-280, 1995.

[2] S.E.Robertoson, K. Sparck Jones: *"Relevance Weighting of Search Terms"*, Journal of the American Society for Information Science, 27, pp.129-146, 1976.

[3] K.L.Kwok: *"Experiments with Component Theory of Probabilistic Information Retrieval Based on Single Terms as Document Components"*, ACM Trnasactions on Information Systems, 8(4), pp.363-386, 1990.

[4] N.Fuhr: *"Models for Retrieval with Probabilistic Indexing"*, Information Processing and Retrieval, 25(1), pp.55-72, 1989.

[5] M. Iwayama, T. Tokunaga: *"A Probabilistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values"*, Proceedings of 4th Conference on Applied Natural Language Processing, pp.162-167, 1994.

[6] M. Iwayama, T. Tokunaga: *"Hierarchical Bayesian Clustering for Automatic Text Classification"*, Proceedings of IJCAI-95, pp.1322-1327, 1995.

[7] K. Aoki, K. Matsumoto, K. Hoashi, K. Hashimoto: *"A Study of Bayesian Clustering of a Document Set Based on GA"*, Proc. of The Second Asia-Pacific Conference on Simulated Evolution And Learning (SEAL98), pp. 260-267, 1998.

[8] M. Iwayama, T. Tokunaga et al.: *"Large-Scale Clustering for Document Search"*, 3rd Annual Meeting of Institute of Language Processing of Japan, pp.245-248, 1997. (In Japanese) Proceedings of SIGIR-95, pp.273-280, 1995.

[9] Goldberg, D.E.: *"Genetic Algorithms"*, Search, Optimization, and Machine Learning, Addison-Wesley, 1989.