類似検索を応用した特許通知システムの試作

2U－2

青木 圭子　松本 一則　帆足 啓一郎　橋本 和夫
KDD 研究所

1. はじめに

筆者らは，これまでに文書中の語の出現確率を用い，文書集合をベイジアンクラスタリングする手法[1]の計算量を削減するため，部分クラスタの評価にMDL基準を用い，準最適なクラスタを遺伝アルゴリズム (以下，GA)[3]を用いて求めることにより高速化する手法[2]を提案した．その結果，大量文書への適用が可能となり，精度も保たれることが分かった．

そこで，大規模な特許文書に適用した特許通知システムを試作した．本稿では，本システムのの概要について報告する．

2. システム要件

従来の特許検索では，一般的に，国際特許番号 (IPC) とキーワードを指定した機械検索の後，人手による絞り込み検索が行われている．しかし，この方法では，

- 検索のための専門知識が必要
- 機械検索だけでは関係のない特許が多く含まれる
- 常に最新情報を得ることが難しい

という問題がある．

これらの問題を解決するため，検索のための専門知識を必要とせず，膨大な特許情報の中から必要な特許のみを検索し，常に最新の情報をユーザに提供することを目的として，以下の2点をシステム要件とした．

- ユーザの興味のある特許文書そのものを入力とする類似特許検索機能
- ユーザ毎に登録された登録された特許を参照して検索を行い，検索結果をユーザへ自動通知する機能

3. 開発システムの概要

3.1 システム構成と動作概要

特許通知システムの構成を図1に示す．

特許データはISDN回線を利用して，定期的に自動送信される．受信特許ファイルは，特許通知サーバにFTP転送され，サーバは受信特許ファイルから特許データを抽出し，新着特許DBを生成する．前月以前の特許については，定期的に過去分特許DBに追加し，クラスタリングを行い，クラスタDBとして保存する．また，ユーザの検索条件をユーザDBに登録しておく．特許通知システムは，新たに特許データを受信した際，ユーザの登録した特許に類似した特許を検索し，発見した場合にはユーザにメイルで通知する．検索結果の詳細はブラウザで閲覧出来る．

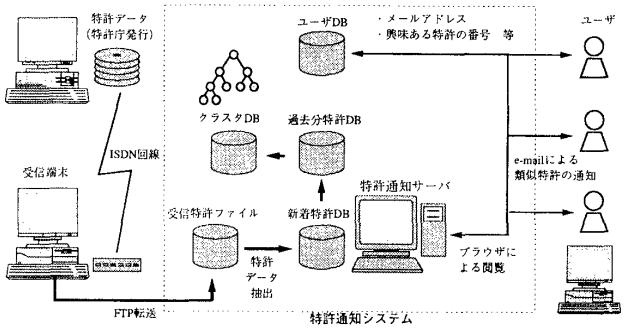

図 1: 特許通知システムの構成

当月の新着特許については，毎回1,000件程度のデータを受信するため，精度の高い総当たり検索を受信日毎に行う．過去に受信した特許については，3ヶ月で約15,000件の公開広報が存在する．そこで高速化のため，クラスタ構築を行った後，クラスタ検索を行う．

3.2 検索条件

検索条件となるユーザDBの登録内容は，以下の通りである．

- 氏名，所属，メイルアドレス
- 登録特許
  発明者に本人の含まれる特許で93年以降の公開公報に含まれる特許が予め登録されている．新たに受信したデータに発明者に本人の含まれる特許が含まれていた場合，自動的に追加登録される．本人以外が発明者となっている特許でも登録可能である．
- 最大検索件数
  デフォルトは各登録特許について，最大10件まで表示する．最大検索件数は登録特許毎に設定可能である．

An Experimental Development of a Selective Dessemination System of Patent Information using Bayesian Clustering
Keiko Aoki, Kazunori Matsumoto, Keiichiro Hoashi, Kazuo Hashimoto
KDD R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka, Saitama 356-8502, Japan

- 類似度の閾値 (ζ)
  デフォルトは各登録特許について，-1.2 の類似度となっている．閾値を上げると検索件数は減り，下げると検索件数は増える．閾値は登録特許毎に設定可能である．

3.3 利用形態

ユーザは検索条件を予め登録しておく．システムが新たに特許データを受信した際，受信したデータ中に類似する特許が含まれている場合には，検索結果が電子メイルに添付されて送付される．図 2に検索結果の例を示す．

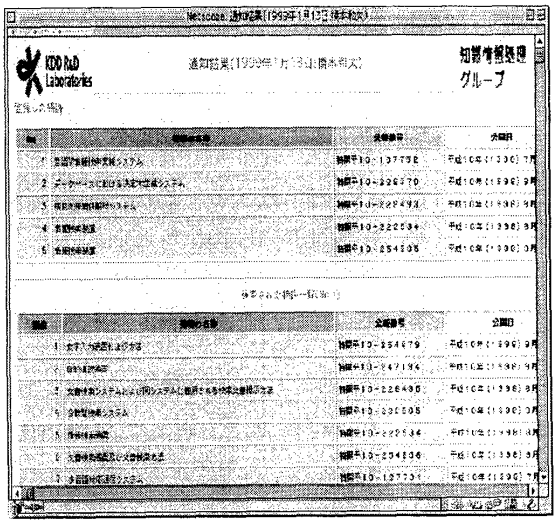

図 2: 検索結果の例

添付ファイルは HTML 形式であり，各データは発明の名称をクリックすることにより内容を閲覧可能である．

4. フィルタリング性能の評価

本システムを用いて，1993 年〜1998 年の間に公開された 21 件の公開広報について同時期に公開された 10,000 件の公開広報から類似する特許の検索を行った．人手による検索結果を正解データとして正解データ，誤りデータの類似度の分布を調べたところ，図 3の様になった．横軸は類似度を表し，値の大きい程，類似度が高いことを示す．縦軸は各類似度における正解，誤りデータの相対頻度であり，1 検索あたりの正解データは平均 10.4 件，誤りデータは平均 9998.6 件である．

誤りデータの殆んどは類似度-1.1 以下に分布しており，誤りデータの類似度- 1.1 以下である確率は 99.7%である．このことから，登録特許毎の閾値を適切に設定することにより，ユーザの求める誤りデータ排除機能を提供できる．

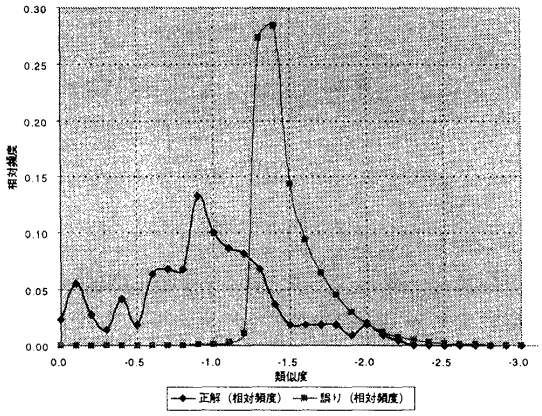

図 3: 正解，誤りデータの頻度分布

5. おわりに

特許通知システムの試作について報告した．本システムを利用することにより，IPC やキーワードの設定等の特許検索のための専門的な知識なしに特許検索が行える様になる．また，自動的に最新の検索結果を送信することにより，最新情報を得ることが出来ると考えられる．今後は，本システムの試行サービスを行い，発明者による類似性の評価を用いたシステム評価を行う予定である．

参考文献

[1] Makoto IWAYAMA, Takenobu TOKUNAGA, "Hierarchical Bayesian Clustering for Automatic Text Classification", Proceedings of IJCAI-95, pp.1322-1327, 1995.

[2] 青木, 松本, 帆足, 橋本, "GA を用いた文書のベイジアンクラスタリングの高速化", 電子情報通信学会技術研究報告, AI98-14, pp.99-105, 1998.

[3] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley(1989).