多言語に特化した特許検索システム（仮称 atari-kun）の構築

亀谷　展

株式会社サン・フレア　自然言語処理技術部
〒160-0004 東京都新宿区四谷 4–7 新宿ヒロセビル
E-mail: kameya_h@sunflare.co.jp

A Multi-language Patent Retrieval System (Atari-kun)

Hiroshi KAMEYA

Natural Language Processing Technology, SunFlare Co., Ltd.
Shinjuku Hirose Bldg., 4-7, Yotsuya, Shinjuku-ku, Tokyo 160-0004
E-mail: kameya_h@sunflare.co.jp

概要

厳しい景気後退とその後に続く景気停滞を受けた国内企業は、生産活動と販売活動の比重を海外に移しつつある。この流れに従い増加傾向にある海外特許出願には出願先国での先行技術調査が不可欠であり、その手間とコストは無視できないものとなっている。例えば、海外の調査会社や特許事務所に調査を依頼すれば多額の費用がかかる上に、これら依頼先とのやり取りは相手国言語もしくは英語を介して行わなければならず煩雑な作業を強いられる。その上、その調査結果が妥当であるかどうかも調査結果の翻訳と内容の吟味を必要とし、調査結果の有効活用には何らかの支援が必要である。本研究では、効果的な先行技術調査の支援を目的とし、弊社の多言語リソースと自然言語処理技術を組み合わせ、企業内技術者の先行技術調査の支援を行う特許検索システムを構築した。本システムは、日本語←→多言語間を翻訳するモジュールと、明細書から IPC を推定するモジュールと、明細書からその特徴をよく示す重要語を抽出するモジュールと、特許明細書を検索するモジュールと、それらの結果をレイアウトするモジュールとで構成される。世界知的所有権機関のサイトで配布されているテキスト分類コーパス WIPO-alpha により本システムの IPC 推定モジュールを評価したところ、Fall ら [25] の報告と同等かそれ以上の精度を得た。

1　はじめに

研究開発により新たに生み出された技術やビジネスは、特許権を取得することで独占的に利用することができる。しかしそれらに新規性と進歩性が無ければ特許として認められず、かけた時間と費用が無駄になってしまう。重複投資や重複研究を避けるためには研究開発に先駆けて先行技術を調査することが望ましい。特に、景気後退とその後に続く景気停滞を受けて国内企業が生産活動と販売活動の比重を海外に移そうとするなら、海外での特許出願と出願先国での先行技術調査が不可欠となる。このような先行技術調査は特許事務所や特許調査会社に依頼したり自分で調査する方法が考えられるが、いずれにしても膨大な特許データを言語の壁を越えて検索するシステムが必要である。本稿では、こうした背景から開発中である多言語に特化した特許検索システム（仮称 atari-kun）について紹介する。

ドキュメント集合から希望するドキュメントを取り出すタスクは情報検索と言われ古くから研究されている [1, 2, 3]。例えば Salton らは SMART (System for the Mechanical Analysis and Retrieval of

Text）と呼ばれる検索システムを開発した [1, 3]。彼らの検索システムは、ドキュメントを構成する単語集合から不要語（ストップワード）リストに含まれる単語を除いた単語集合でドキュメントを表す多次元ベクトルを求め、同様に検索質問（クエリ）からも多次元ベクトル表現を求め、それらのベクトルから類似度を計算することによってドキュメント検索を実現するベクトル空間モデルに基づいている。ドキュメントを構成する単語集合からベクトル表現を求めるとき、各々の単語すべてが同じようにドキュメントの内容に関係するわけではないので、単語の持つ重要度に応じて重み付けがなされることが多い。例えば、冠詞や助動詞といった機能語がほぼ全てのドキュメントに現れるのに対し、名詞などの内容語はドキュメントの内容に大きく関係して現れるので、ベクトル表現において内容語は機能語よりも大きな重み付けをした方がドキュメントの内容をより正確に表現することができ検索精度も向上する [4, 5]。このように多くのドキュメントに現れる語の重みを小さく特定のドキュメントに偏って現れる語の重みを大きくするように単語の頻度（term frequency; TF）とドキュメント頻度の逆数（inverse document frequency; IDF）を用いてドキュメント内での単語の重要度を計算する方法はTF-IDF[1, 3] と呼ばれ、ベクトル空間モデルに基づく情報検索システムにおける基本的な重み付け方法として知られる。IDF は同じドキュメント頻度でもドキュメント毎の出現頻度が異なる語を区別しないため、内容語と単なる数字や一般的に現れる語（一般語）に異なる重みをつけることができないという欠点を持つ。これに対して、ポアソン分布が一般語に対してあてはまりそれ以外の内容語に対してあてはまらないという観点から、ドキュメントにおける単語の出現頻度にポアソン分布をあてはめ推定したドキュメント頻度と実測値の差を単語の重要度とする残差 IDF（residual IDF; RIDF）が Church と Gale により提案された [6, 7]。特許検索に特化した重要語の選択法としては、小西ら [8] が先願特許検索の検索語を抽出するためのヒューリスティックなルールを提案している。しかし多言語対応を視野に入れた特許検索システムでは言語の数だけルールを人手で定義しなければならず大きなコストがかかってしまうので、ドキュメント集合が存在すれば数学的な手続きで語の重要度を決定できる RIDF は多言語に特化した特許検索システムに適していると思われる。

日本語で書かれたクエリで他言語のドキュメント集合を検索するタスク、すなわち記述言語がクエリとドキュメントで異なる情報検索タスクは言語横断検索（Cross-Language Information Retrieval; CLIR）と呼ばれる [9, 10, 11]。言語横断検索システムは次の 3 つに分類できる [12]。

1. ドキュメント集合をクエリの言語に翻訳する。
2. クエリをドキュメント集合の言語に翻訳する。
3. クエリとドキュメント集合を別の表現に変換する。

1 では検索に先立って翻訳しておかねばならないため、検索システム開発時に時間的・金銭的コストが多くかかる。そのため人手翻訳ではなく機械翻訳で済ませたり、検索範囲を特定の範囲に限定してしまうことが多い。例えば市販で提供されている特許検索システムの多くはこのような方法で日本語以外の特許明細書の検索に対応している。2 では検索時に動的にクエリを翻訳する。例えば Fujii と Ishikawa[13] はタスクを 2 段階に分け第 1 ステージでクエリを翻訳して検索し、第 2 ステージで検索結果のドキュメントを翻訳しクラスタリングして出力することを提案した。Higuchi ら [14] は Fujii と Ishikawa のアイデアに基づくシステムにパテントファミリーから複合語の対訳を自動抽出するモジュールの追加を試みその有効性を示した。3 ではクエリとドキュメント集合をそれらとは独立した表現に変換する。例えば Salton[15] はシソーラスを用いてクエリとドキュメント集合の言語の違いを解消することを提案した。また Littman ら [16] はパラレルコーパスから計算した潜在的意味インデックス（Latent Semantic Indexing）を用いても可能なことを示した。しかし 1 と同様にコストの問題が残る。これらの検索精度について、Oard[17] や

McCarley[18] は、言語横断検索においては翻訳方向の違いによる翻訳精度の影響が大きいことを指摘している。

以上を踏まえ、atari-kun では上記 2 の手法を選び、クエリをなるべく高精度に翻訳することで検索精度を高めることを目指した。特許検索用のクエリは特許の特徴を表す技術用語で構成される。つまり検索精度を高めるにはクエリ内の技術用語を特許の内容に応じて適切に訳しわけることが必要であると思われる。特許には国際特許分類（IPC）と呼ばれる世界 90 ヶ国以上の国で利用されている分類がある [19]。IPC は特許の技術内容により分類されているので、IPC 毎に専門用語辞書を作成しそれらを機械翻訳に用いることで翻訳の精度を向上させられることが期待できる。

以下、2 節では開発中のシステムについて述べる。3 節で本システムのモジュールの実験結果を示す。4 節でまとめと今後の課題について述べる。

2 多言語に特化した特許検索システム atari-kun の開発

2.1 システムの構成

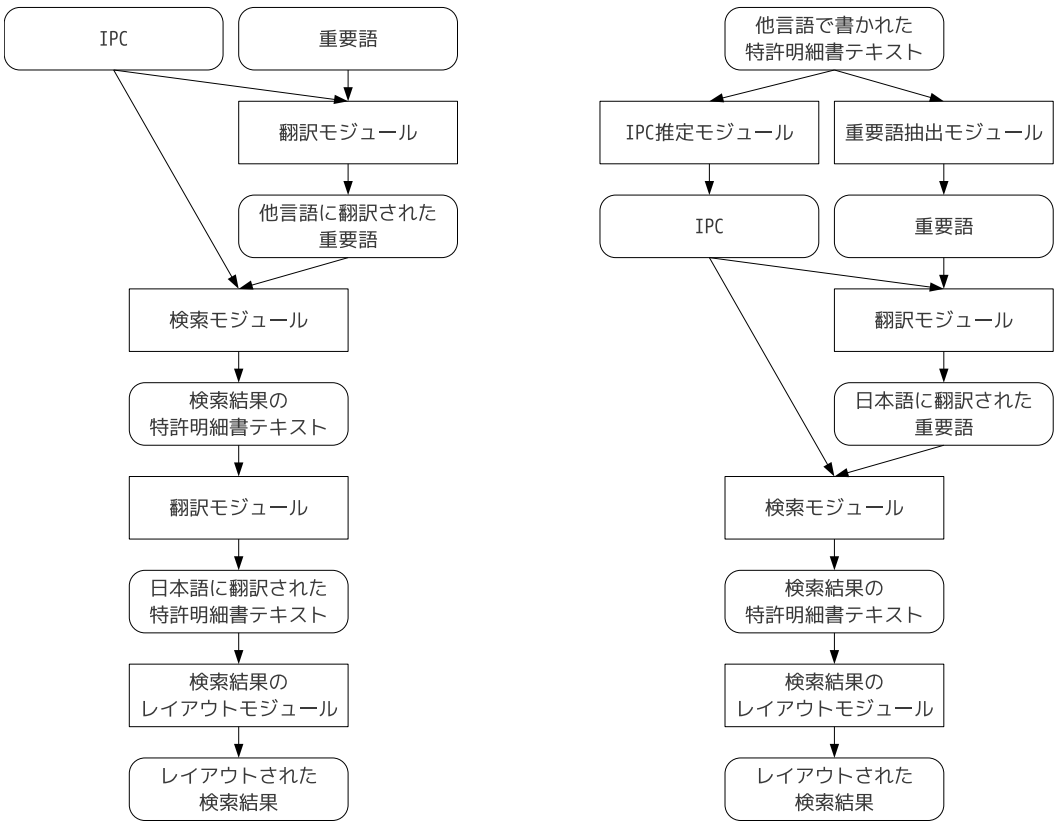
図 1 にシステムの概要を示す。図 1(a) は内外出願（日本から海外への出願）、同図 (b) は外内出願（海外から日本への出願）向け検索の処理を図示している。

内外出願の場合、特許明細書は日本語で書かれているのでユーザーは自ら重要語を選択することができるし、出願済みであれば IPC も判明している。従って atari-kun には重要語やドキュメントを翻訳するモジュールと IPC と重要語を元に検索するモジュールと検索結果をレイアウトするモジュールが有れば良い。

一方、外内出願の場合、他国語が分からなければ明細書を読むことができないため重要語を選択することもできない。IPC も出願時の IPC しか分からず、中身を理解した上で IPC を追加して検索するということもできない。そこで、atari-kun は外内出願向け検索のために明細書からその特徴をよく示す重要語を抽出するモジュールと明細書から IPC を推定するモジュールも含んでいる。

以下、各モジュールについて説明する。

- 翻訳モジュール
  翻訳モジュールはルールベース型機械翻訳エンジンと IPC のセクション毎に作成された専門用語辞書で構成される。
- 重要語抽出モジュール
  重要語モジュールは IPC の全分野を網羅するように収集した特許コーパスから学習した重要語判定エンジンにより重要語を抽出する。
- IPC 推定ジュール
  文書に 1 つ以上のタグを付与するタスクを文書分類 [1, 3] と呼ぶ。IPC の推定も活発に研究されており様々なアルゴリズムが提案されている [20, 21, 22, 23, 24]。Fall ら [25] は 3 つの代表的な学習器（k-NN と Naive Bayes と SVM）による実験の結果を報告している。Naive Bayes による文書分類アルゴリズムは、文書に含まれる単語の出現確率が他の単語とは独立であるというシンプルな仮定に基づいており、SVM や非線形分類器による複雑なアルゴリズムに比べて実装が容易で実行時間が少ないというメリットを持つ。そこで IPC 推定モジュールは Naive Bayes を用いてアルゴリズムを構築した。Fall らの Naive Bayes を用いた実験ではドキュメントベクトルの要素に単語の出現確率をスムージングしたものを使用しているが、本研究では単語の重要度をドキュメントベクトルの要素とした。そして IPC の全分野を網羅するように収集した特許コーパスで学習させた。
- 検索モジュール
  検索モジュールは IPC と重要語から自動的に生成した検索クエリで特許データベースを検索する。
- レイアウトモジュール
  レイアウトモジュールはクライアントのカスタマイズに対応できるように構成される。



(a) 内外出願向け検索　　(b) 外内出願向け検索

図 1　atari-kun のシステム

2.2　システムの機能

ユーザーが atari-kun で検索を行うには、検索条件などを設定した検索ジョブを作成する。検索の結果は atari-kun がレポートにまとめてユーザーに返してくる、という流れになっている。

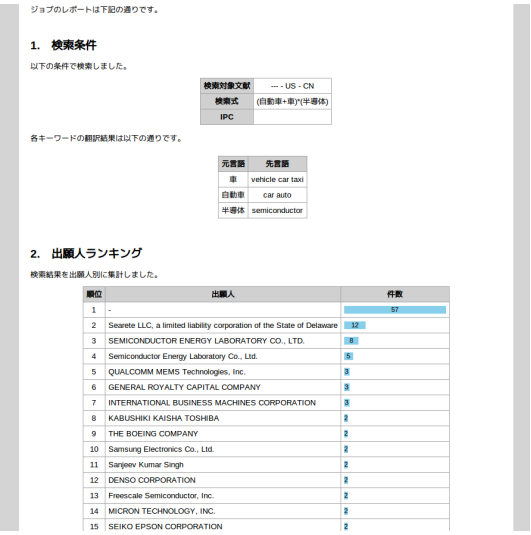
まず、図 2 に検索ジョブの新規作成画面を示す。

図 2　検索ジョブ作成画面の例

1. 内外出願向け検索
   内外出願向け検索を行う場合、ユーザーは画面上部にあるフォームを操作して検索ジョブを atari-kun に登録する。まず検索対象文献を指定し、次に検索式を日本語で入力する。次に IPC を指定し、Create ボタンをクリックするとジョブの作成が完了する。
2. 外内出願向け検索
   外内出願向け検索を行う場合、ユーザーは画面下部にあるフォームを操作して検索ジョブを atari-kun に登録する。まず doc または pdf ファイルを指定するか、他言語で書かれた明細書テキストをテキストボックスにペーストする。次に Create ボタンをクリックするとジョ

ブの作成が完了する。

検索ジョブが完了するとユーザーにはメールが届き、画面上で検索結果を確認できるようになる。図 3 に検索ジョブを実行した結果の画面の例を示す。



ジョブのレポートは下記の通りです。

1. 検索条件

以下の条件で検索しました。

検索対象文献	— - US - CN
検索式	(自動車+車)^(半導体)
IPC	

各キーワードの翻訳結果は以下の通りです。

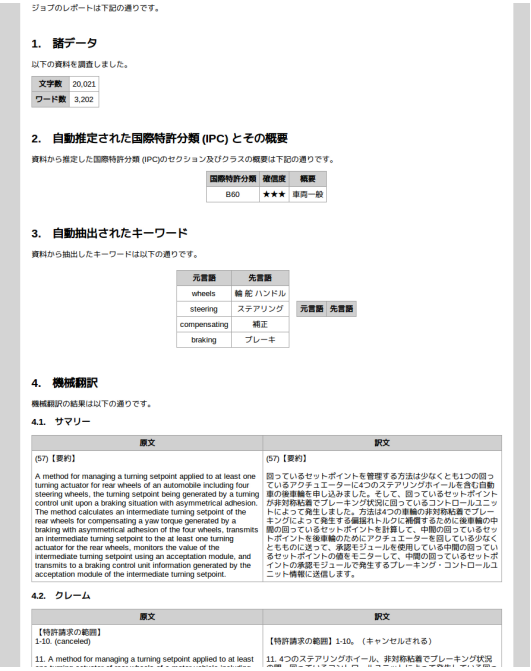
元言語	先言語
車	vehicle car taxi
自動車	car auto
半導体	semiconductor

2. 出願人ランキング

検索結果を出願人別に集計しました。

順位	出願人	件数
1	-	57
2	Searete LLC, a limited liability corporation of the State of Delaware	12
3	SEMICONDUCTOR ENERGY LABORATORY CO., LTD.	8
4	Semiconductor Energy Laboratory Co., Ltd.	5
5	QUALCOMM MEMS Technologies, Inc.	3
6	GENERAL ROYALTY CAPITAL COMPANY	3
7	INTERNATIONAL BUSINESS MACHINES CORPORATION	3
8	KABUSHIKI KAISHA TOSHIBA	2
9	THE BOEING COMPANY	2
10	Samsung Electronics Co., Ltd.	2
11	Sanjeev Kumar Singh	2
12	DENSO CORPORATION	2
13	Freescale Semiconductor, Inc.	2
14	MICRON TECHNOLOGY, INC.	2
15	SEIKO EPSON CORPORATION	2

(a) 内外出願向け検索の結果の一部



ジョブのレポートは下記の通りです。

1. 諸データ

以下の資料を調査しました。

文字数	20,021
ワード数	3,202

2. 自動推定された国際特許分類 (IPC) とその概要

資料から推定した国際特許分類 (IPC)のセクション及びクラスの概要は下記の通りです。

国際特許分類	確信度	概要
B60	★★★	車両一般

3. 自動抽出されたキーワード

資料から抽出したキーワードは以下の通りです。

元言語	先言語
wheels	輪 舵 ハンドル
steering	ステアリング
compensating	補正
braking	ブレーキ

元言語	先言語

4. 機械翻訳

機械翻訳の結果は以下の通りです。

4.1. サマリー

原文	訳文
(57)【要約】 A method for managing a turning setpoint applied to at least one turning actuator for rear wheels of an automobile including four steering wheels, the turning setpoint being generated by a turning control unit upon a braking situation with asymmetrical adhesion. The method calculates an intermediate turning setpoint of the rear wheels for compensating a yaw torque generated by a braking with asymmetrical adhesion of the four wheels, transmits an intermediate turning setpoint to the at least one turning actuator for the rear wheels, monitors the value of the intermediate turning setpoint using an acceptation module, and transmits to a braking control unit information generated by the acceptation module of the intermediate turning setpoint.	(57)【要約】 回っているセットポイントを管理する方法は少なくとも1つの回っているアクチュエーターに4つのステアリングホイールを含む自動車の後車輪を申し込みました。そして、回っているセットポイントが非対称粘着でブレーキング状況に回っているコントロールユニットによって発生しました。方法は4つの車輪の非対称粘着でブレーキングによって発生する偏揺れトルクに補償するために後車輪の中間の回っているセットポイントを計算して、中間の回っているセットポイントを後車輪のためにアクチュエーターを回している少なくともセットポイントの値をモニターして、中間の回っているセットポイントの承認モジュールで発生するブレーキング・コントロールユニット情報に送信します。

4.2. クレーム

原文	訳文
【特許請求の範囲】 1-10. (canceled)	【特許請求の範囲】1-10。（キャンセルされる）
11. A method for managing a turning setpoint applied to at least	11. 4つのステアリングホイール、非対称粘着でブレーキング状況

(b) 外内出願向け検索の結果の一部

図 3　検索結果画面の例

3　実験

atari-kun は開発中の段階にあり、多言語対応では英語・中国語・ドイツ語・ポルトガル語・韓国語の 5 言語を始めとして更に対応言語を増やしていく予定だが、全てのモジュールが十分なレベルで機能するほど完成しているというわけではない。そこでこの節では IPC 推定モジュールの実験結果を示す。

実験には世界知的所有権機関（WIPO）のサイトで配布されているテキスト分類コーパス WIPO-alpha[26] を用いた。WIPO-alpha は英文特許明細書のコーパスで、約 46,000 件（約 3 億ワード）の訓練セットと、約 29,000 件（約 2.4 億ワード）のテストセットで構成されている。

IPC 推定モジュールを WIPO-alpha の訓練セットで学習させ、テストセットで分類させた結果を表 1 に示す。分類のレベルはクラスである。表内には、比較として Fall ら [25] の報告から分類用のテキストとして明細書のタイトル・発明人・出願人・アブストラクト・詳細な説明の最初の 300 語を用いて Naive Bayes で分類した場合の結果も示した。特許明細書には複数の IPC が付与されることもある。そこで推定精度の評価方法としては Fall ら [25] の top prediction、three guesses、all categories を採用した。top prediction では分類器の出力の第 1 位の IPC が明細書のメイン IPC と一致したときに正解とカウントする。three guesses では分類器の出力の第 1 位〜第 3 位のどれかが明細書のメイン IPC と一致したときに正解とカウントする。all categories では分類器の出力の第 1 位の IPC が明細書に付与されたどれかの IPC と一致したときに正解とカウントする。

atari-kun に IPC 推定モジュールの推定精度は、どの評価法であっても Fall らのものより優れていることが確認できた。特に three guesses では 86.8% もの高精度で推定できていることから、外内出願向け検索で IPC 推定モジュールの上位 3 位までを検索条件に盛り込むことにより検索結果の精度向上に繋げられると期待できる。

また、セクションレベルの top prediction では

表 1　IPC 推定の結果

手法	Top-prediction	Three-guesses	All-categories
Fall ら [25]	55%	79%	63%
atari-kun	**62.7%**	**86.8%**	**70.4%**

73.3% の精度だった。

4　おわりに

多言語に特化した特許検索システム（仮称 atari-kun）を紹介した。

まず国内企業の特許を取り巻く状況から多言語に特化した特許検索システムに必要な要件について考察し、それに基づくシステムの概要を示した。

システムは開発途中の状態ではあるが、システムを構成するモジュールのうち IPC 推定モジュールを WIPO-alpha にて評価し、Fall らの報告した精度より高い精度で IPC 推定できたことを確認できた。

今後は、内外出願向け検索ジョブの IPC とクエリを与えたときの翻訳精度と検索精度の検証と、外内出願向け検索ジョブの重要語の抽出精度と検索精度の検証を行う予定である。また、多言語対応としてコーパス作成と辞書構築の完成を急ぎ、多言語に特化した特許検索システムをできるだけ早く提供することを目指す。

参考文献

[1] Christopher D. Manning, Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[2] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002.

[3] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[4] Karen S. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[5] Gerard Salton, Christopher Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, vol. 24, issue 5, pp. 513–523, 1988.

[6] Kenneth W. Church, William A. Gale. Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. In *Proceddings of the Third Workshop on Very Large Corpora*, pp. 121–130, 1995.

[7] Kenneth W. Church, William A. Gale. Poisson Mixtures. *Natural Language Engineering*, vol. 1, pp. 163–190, 1995.

[8] 小西一也, 北内啓, 高木徹. 発明の特徴に着目した検索語抽出による先願特許検索. 第 15 回データ工学ワークショップ DEWS2004 3-B-01.

[9] Douglas W. Oard, Bonnie J. Dorr. A Survey of Multilingual Text Retrieval. *Technical Report UMIACS-TR-96-19*, University of Maryland, Institute for Advanced Computer Studies, 1996.

[10] Douglas W. Oard. Alternative Approaches for Cross-Language Text Retrieval. In *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, pp. 131–139, 1997.

[11] 樋口重人, 福井雅敏, 藤井敦, 石川哲也. 特許情報を対象とした言語横断検索システム. 言語処理学会第 7 回年次大会発表論文集, pp. 445–447, 2001.

[12] Atsushi Fujii, Tetsuya Ishikawa.

Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration. *Computers and the Humanities*, vol. 35, pp. 389–420, 2001.

[13] Atsushi Fujii, Tetsuya Ishikawa. Applying Machine Translation to Two-Stage Cross-Language Information Retrieval. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas*, pp. 13–24, 2000.

[14] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, Tetsuya Ishikawa. PRIME: A System for Multi-lingual Patent Retrieval. In *Proceedings of MT Summit VIII*, pp. 163–167, 2001.

[15] Gerard Salton. Automatic Processing of Foreign Language Documents. *Journal of the American Society for Information Science*, vol. 21, issue 3, pp. 187–194, 1970.

[16] Thomas K. Landauer, Michael L. Littman. A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments. In *Proceedings of the 11th International Conference on Expert Systems and their Applications*, vol. 8, pp. 77–85, 1991.

[17] Douglas W. Oard. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americans on Machine Translation and the Information Soup*, pp. 472–483, 1998.

[18] J. Scott McCarley. Should we Translate the Documents or the Queries in Cross-language Information Retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 208–214, 1999.

[19] 高柳隆. 「特許調査」の基礎と応用. 工学図書, 2006.

[20] Caspar J. Fall, K. Benzineb. Literature survey: Issues to be Considered in the Automatic Classification of Patents. *World Intellectual Property Organization*, *http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/wipo-categorizationsurvey.pdf*, 2002.

[21] Shantanu Godbole, Sunita Sarawagi. Discriminative Methods for Multi-Labeled Classification. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 22–30, 2004.

[22] Lijuan Cai, Thomas Hofmann. Hierarchical Document Categorization with Support Vector Machines. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pp. 78–87, 2004.

[23] Juho Rousu, Craig Saunders, Sandor Szemak, John Shawe-Taylor. Learning Hierarchical Multi-Category Text Classification Models. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 744-751, 2005.

[24] Akinori Fujino, Hideki Isozaki, Jun Suzuki. Multi-label Text Categorization with Model Combination based on F_1-score Maximization. In *Proceedings of the 3rd International Joint Conference on Natural Language Porcessing*, pp. 823-828, 2008.

[25] Caspar J. Fall, A. Törcsvári, K. Benzineb, G. Karetka. *Automated Categorization in the International Patent Classification.* ACM SIGIR Forum, vol. 32, issue. 1, pp. 10–25, 2003.

[26] WIPO-alpha. *http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html*