

タキソノミーを用いた Web 統合検索処理の効率化

崔 春花[†] サイド ミルザ パレビ[‡] 北川 博之^{*}

あらまし インターネット技術の発展に伴い、Web 情報の高精度の検索の必要性が高まっている。Web 情報検索の精度を向上するための手段の一つとして、タキソノミーの利用がある。我々の研究グループでは、タキソノミーを有するディレクトリ型検索エンジンを併用することで、Web 上にある他の情報源の検索精度を向上することが可能な Web 統合検索処理手法を提案した。本手法では、利用者は検索キーワードに加えて、ディレクトリ型検索エンジンが提供するタキソノミー中の適当なノードをコンテキストノードとして指定する。システムはこの情報を用いた適応型の問合せ拡張を行うことで、検索コンテキストにより適合した問合せを対象情報源に対して行うことができる。しかし、本手法ではディレクトリ型検索エンジンから問合せ処理時に動的に情報を獲得するため、この処理時間が問合せ実行時間に大きく影響を及ぼす場合があるという問題がある。本研究では、動的に獲得する情報の量を大幅に削減する改良手法を提案し、実際の Web 検索サイトを用いた実験により改良手法の有効性を評価する。

キーワード Web 検索, タキソノミー, 適応型問合せ拡張

Performance Improvement of Taxonomy-based Integrated Web Retrieval

CHUNHUA CUI[†] SAID MIRZA PAHLEVI[‡] and HIROYUKI KITAGAWA^{*}

Abstract As the Internet technology evolves, the need for high web retrieval effectiveness continues to escalate. One way to improve web retrieval effectiveness is by using taxonomy. Our research group has proposed a taxonomy-based integrated web retrieval method that utilizes taxonomy-based search engines such as web directories to improve the search effectiveness of the other web search engines. In the proposed method, a user specifies a context category from taxonomy of a web directory search engine besides providing search keywords. System then adaptively modifies the user query using the given information so that the query becomes more focus on the selected context category (search context). However, since the proposed method adaptively probes the web directory for retrieving needed information, the query response time might become slightly longer. In this paper, we extend the proposed method by decreasing the amount of information retrieved from the web directory so that the query processing time can be shortened. We evaluate the effectiveness of the extended method by using real web search engines.

Keyword Web Retrieval, Taxonomy, Adaptive Query Expansion

[†]筑波大学理工学研究科

Master's Program in Science and Engineering, University of Tsukuba

[‡]産業技術総合研究所

National Institute of Advanced Industrial Science and Technology

^{*}筑波大学電子・情報工学系

Institute of Information Sciences and Electronics, University of Tsukuba

1. はじめに

インターネット技術の発展に伴い、Web 情報の高精度の検索の必要性が高まっている。広い範囲で利用される検索エンジンとして、クローラに基づく検索エンジンとタキソノミーに基づく検索エンジンが知られている。

例えば、Google はクローラに基づく検索エンジンの代表的な例である。クローラに基づく検索エンジンは、人間のわずかの介在だけでウェブ・ページを自動的に辿って行くため、比較的広い範囲の Web 情報をカバーすることが可能である。しかし、検索エンジンが取ってきたページを、そのトピックとは無関係にキーワードの集合とみなして検索の主な手がかりとして用いる。したがって、適切なキーワードが与えられない場合には、検索結果は一般に多くのノイズを含み得る。また、適切なキーワードを与えること自身も難しい場面も多い。一方、タキソノミーに基づく検索エンジンでは概念的階層を使用して複雑さを管理する。例としては、Yahoo、ODP などがある。これらの検索エンジンでは、利用者は検索時にキーワードを与えるだけでなく検索のコンテキストを表すカテゴリを指定することが可能である。キーワードとコンテキストカテゴリを組み合わせることで情報の絞り込みを行うため、一般に検索結果の適合率を高めることが可能である。しかし、これらの検索エンジンで検索対象とできるのは、それが内包するデータベース中であらかじめ分類されている Web 情報のみである。また、分類は人手で行われるため、検索可能な情報の範囲はクローラに基づく検索エンジンに比べて非常に小さい。

我々の研究グループでは、既存のクローラ型検索エンジンとタキソノミーを有するディレクトリ型検索エンジンを併用することで、Web 上にある他の情報源の検索精度を向上することが可能な Web 統合検索処理手法について研究を行ってきた[1, 2]。本手法では、利用者は検索キーワードに加えて、ディレクトリ型検索エンジンが提供するタキソノミー中の適当なノードをコンテキストノードとして指定する。システムはこの情報を用いた適応型の問合せ拡張を行うことで、検索コンテキストにより適合した問合せを対象情報源に対して行うことができる。

しかし、本手法ではディレクトリ型検索エンジンから問合せ処理時に動的に情報を獲得することが必要なため、その処理時間が問合せ実行時間に大きく影響を及ぼす場合があるという問題がある。本研究では、ディレクトリ型検索エンジンより動的に獲得する情報の量を削減する改良手法を提案する。また、実際の Web 検索サイトを用いた実験により改良手法の有効性を評価する。

本論文の構成は以下の通りである。2 節では我々の従来の提案手法についてより詳しく説明する。3 節では、本論文で提案する 2 種類の改良手法について記述する。4 節では提案手法の評価実験および結果を示す。5 節ではまとめと今後の課題について述べる。

2. 従来提案の統合検索方式

我々は Web 検索効率の向上のため、ディレクトリ型検索エンジンが提供するタキソノミーを利用する統合検索手法を提案した。この手法では、通常のディレクトリ型検索エンジンにおける場合と同様に、利用者は問合せ要求を与えるために、検索キーワードとタキソノミー中のコンテキストカテゴリを指定する。システムは、ディレクトリ型検索エンジンからその検索キーワードを含むページのうち、コンテキストカテゴリに合致するものとそうでないものを抽出する。さらに、システムは前者を正例、後者を負例として学習を行うことで、コンテキストカテゴリに合致するページを特徴づけるキーワードを得る。このキーワードを付加して利用者問合せを拡張し、その拡張問合せをクローラ型検索エンジン等の対象検索エンジンに送付する。これにより、ディレクトリ型検索エンジン以外の情報源に対しても、タキソノミーを用いた精度の良い検索を可能とする。図 1 はこのような統合検索方式の流れを示している。以下では、これらの手順についてより詳しく説明する。

2.1 問合せの定式化

問合せの定式化の手続きはタキソノミーを提供するディレクトリ型検索エンジンにおける場合と同様である。適切な情報を最初に見つけるために、利用者はタキソノミーの分類階層を辿って

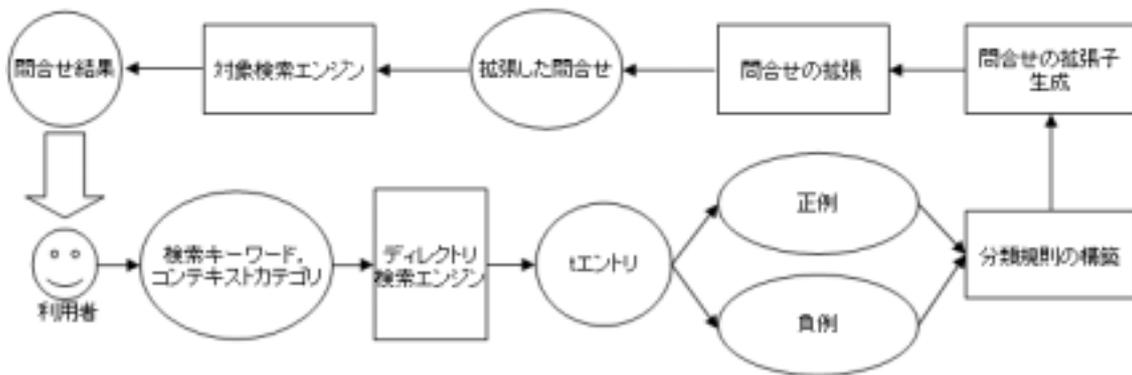


図1 従来の提案手法の流れ

いく。そして、探したいトピックにふさわしいカテゴリ G を選択した後、検索キーワード Q に基づいた問合せ (Q, G) を構築する。以下では、利用者によって選択されたカテゴリをコンテキストカテゴリと呼ぶ。

2.2 ディレクトリ型検索エンジン情報のプロービング

利用者からの問合せ (Q, G) に基づき、システムはディレクトリ型検索エンジンから t エントリの取得を行う。ただし、 t エントリとは、ディレクトリ型検索エンジンが保持する検索対象単位のこと、通常は、Web サイトの URL とその説明記述などの組合せである。具体的には、キーワード Q を含み G に分類される t エントリを正例、キーワード Q を含み G に分類されない t エントリを負例として取得する。しかし、これらの t エントリの取得には膨大な時間がかかる。そのため、次の手順で t エントリの取得を行う。

1. 利用者からの問合せ (Q, G) と $(Q, _)$ をディレクトリ検索エンジンに送る。ここで、 $(Q, _)$ はコンテキストカテゴリを指定しないことを意味する。この問合せに対しては、 Q を含むすべての t エントリが解となる。ディレクトリ型検索エンジンの検索結果の先頭ページには通常条件を満たす全 t エントリ数が記載される。したがって、これらの問合せ結果より、 (Q, G) に対する全正例数 P と全負例数 N を得ることができる。
2. $SP = p+q * P / (P+N)$ により実際に取得する正例数を決定し、問合せ (Q, G) の検索結果からそれらを取得する。
3. $SN = p+q * N / (P+N)$ により実際に取得する負例数を決定する。これらの負例はトップレベルにある各カテゴリ G_i 下より平均的に取得する。ただし、利用者が指定したコンテキストカテゴリが G_i のいずれかの場合にはそれを除く。また、コンテキストカテゴリがある G_i のサブカテゴリの場合は、各 t エントリがもつカテゴリ情報を用いて負例のみを分離する。

上記 2, 3 において、 p は正例と負例の最小数を表す、 q は各 p 個の正例、負例に加えて取得する正例および負例の総数である。

2.3 問合せの拡張と実行

上記により取得した正例と負例を学習データとして分類規則を導出する。分類規則は、“H Relevant”の形式をもつ。ただし、Relevant は正例に対応するクラスである。また、H は下記の形式をもつ条件式である。

Google をはじめとする多くのクローラ型検索エンジンはこの形式の条件式を受理可能である。

$$H = (w_{1,1} \wedge \dots \wedge w_{1,i}) \wedge (w_{2,1} \wedge \dots \wedge w_{2,j}) \wedge (w_{3,1} \wedge \dots \wedge w_{3,k})$$

ここで $w_{m,n}$ は検索条件を表す語に相当し、 i, j, k は 0 以上の整数である。ただし、 $k \geq 2$ とする。問合せ拡張を行う際に考慮しなければならないもう一つの制約は問合せサイズである。例えば、Google の場合、問い合わせの最大長は 10 個語である。我々の従来の統合検索方式では、新たな学習アルゴリズムを開発し、これらの点を考慮した分類規則を生成する。

このようにして分類規則“H Relevant”を得た後、システムは H を問合せ拡張子として抽出し、拡張問合せ Q H を対象検索エンジンに送る。最後に、その検索結果を利用者に返す。

3. 提案手法

上記の従来提案手法では、分類規則を導出するための 2.2 節で述べた方法で決定した SP 個の正例と SN 個の負例を、問合せ処理時に動的にディレクトリ型検索エンジンより取得していた。SP と SN の値は拡張問合せの検索精度と関係し、一般に p, q の値を大きくして正例と負例の数を増加した方が検索精度の向上が高い、ことがこれまでの実験で確認されている。しかし、このためにはディレクトリ型検索エンジンと多くのインタラクションを必要とする。このインタラクション時間は、全体の問合せ処理時間を決定する主要な部分であり、SP と SN を増加させた場合には、全体の問合せ処理時間に悪影響を及ぼす。そこで、本研究ではこれらの取得方法を改良することで、検索精度の維持とインタラクション時間の削減の両立を図ることを試みる。

2.2 節に述べた方法では、基本的には、Web 上の正例と負例の比率に応じて SP と SN を決定していた。多くの状況では、 $SP < SN$ となり、特に SN 個の負例を取得する時間がインタラクション時間の中で大きな割合を占める。したがって、本研究では、負例の取得時間を減らす、以下の 2 つの手法を提案する。

手法 1 :

従来手法では、正例と負例の存在比率に応じて SP と SN を決定していたが、SN を SP と同数とする。上に述べたように、従来手法ではほとんどの場合 $SP < SN$ となるので、この方法では従来手法で同数の正例を取得する場合に比べて、實際上取得する負例数 SN を削減することになる。

手法 2 :

あらかじめディレクトリ型検索エンジンが保持する t エントリの一部を抽出し、ローカルな t エントリファイルに格納する。この中からランダムにサンプリングした t エントリを負例として用いる。この場合は、負例の取得にはディレクトリ型検索エンジンとのインタラクションを全く必要としない。このようにして取得する負例の t エントリ数は、本研究では手法 1 と同数とする。

所定の SP 個の正例を取得する場合を考えると、手法 1 では、取得する負例の数が削減される分だけディレクトリ型検索エンジンとのインタラクション時間が削減される。しかし、負例を削減した影響が検索精度に現れる可能性がある。一方、手法 2 では、負例をローカルな t エントリファイルよりランダムに取得する。したがって、負例の取得の際、ローカルファイルアクセスは必要になるもの、ネットワークを介したディレクトリ型検索エンジンとのインタラクション時間全体が削減される。ただし、従来手法では指定されたキーワード Q を含むテキストカテゴリに分類されない t エントリを用いていたのに対し、ランダムにサンプリングした t エントリを用いることになるため、若干の検索精度の低下が予想される。

方法 2 の変形としては、あらかじめローカルに保持する t エントリをディレクトリ型検索エンジンのもつタキソノミーにしたがって分類しておくこと、キーワード Q の有無を考慮したサンプリングを行うことなどが考えられる。しかし、いずれの場合でも、t エントリファイルを保持・検索するためのより複雑な仕組みが必要となる。本研究では、最も単純なランダムなサンプリングを行った場合に焦点を当てて検討を行う。

4. 実験

3 節で述べた方法の有効性を評価するための実験を行う。実験では、問合せ拡張子の導出時間、拡張問合せ結果の適合率、導出された具体的な拡張子を調べ、提案手法 1, 2, 従来手法を比較する。

実験でディレクトリ型検索エンジンとして ODP を、対象検索エンジンとして Google を用いる。提案手法 2 では、予め ODP からランダムに取得した t エントリを格納したファイルを用いる。

実験には、表 1 に示す問合せ 10 個を用いる。それぞれの意味は表 1 に記載の通りであり、これらは ODP のカテゴリ記述に基づく。いずれの問合せにおいても、従来手法では $p=20, q=320$ とし

て正例数 SP と負例数 SN を決定し、提案手法 1, 2 では、従来手法における SP と同一の値を取得する正例数および負例数とする。

4.1 実験 1

実験 1 では、問合せ拡張子の導出時間を実測し比較する。この時間は、ODP に最初に問合せを送付する時点から問合せ拡張子を導出するまでの時間である。提案手法 1 および従来手法では、これは、主に ODP から上記の数の正例と負例を取得し、分類規則を生成するまでの時間である。提案手法 2 では、ODP から負例を取得する部分がローカルファイルから負例を取得することに置き換わる。

図 2 は各問合せに対する 3 手法の問合せ拡張子導出時間のグラフである。各問合せについて、それぞれの手法で 10 回測定した結果の平均時間を示している。このグラフから分かるように、従来手法に比べて、提案手法 1 では相当の時間削減がなされているものの、提案手法 2 ではさらにそれを上回る時間削減がなされている。従来手法と比較した提案手法 1 の導出時間は、最小 48%、最大 84% で、10 個の問合せに対する平均値は 65% である。一方、提案手法 2 については、最小 19%、最大 57% で、10 個の問合せに対する平均値は 32% である。10 個の問合せに関する平均時間を比較したグラフを図 3 に示す。

4.2 実験 2

実験 2 では、拡張問合せを Google に与えた問合せ結果を対象にドキュメントレベル適合率を測定した。カットオフ値 k におけるドキュメントレベル適合率とは、ランキングされた問合せ結果の上位 k 件を対象とした適合率である。本実験では、 $1 \leq k \leq 20$ とし、各 k について 10 個の問合せに対するカットオフ値 k のドキュメントレベル適合率の平均値を計算した。正解か否かの判定は、表 1 に示した各問合せの意味記述に基づき人手で行った。

同一の正例、負例を与えた場合でも、分類規則導出アルゴリズムの内部処理の性質および提案手法 2 でランダムに負例をサンプリングしていることにより、最終的に導出される問合せ拡張子は異なる。そこで、各問合せに関して、20 回問合せ拡張子を導出し、最も頻度が高い問合せ拡張子を適用した場合を対象として測定を行った。図 4 に各カットオフ値 k に対する平均ドキュメントレベル適合率を示す。図 4 では参考のため、拡張を行わない初期問合せに対する同様の適合率も示している。提案手法 1 と従来手法の適合率はほぼ同じである。一方、提案手法 2 では若干適合率の低下が見られるものの、提案手法 1 や従来手法と同様、初期問合せの大幅な改善ができていることが分かる。カットオフ値 $1 \leq k \leq 20$ に関する平均値は、従来手法、提案手法 1、提案手法 2、初期問合せでそれぞれ 88%、87%、77%、28% である。

4.3 実験 3

本実験では、3 手法で導出される問合せ拡張子を比較した。各問合せごとに 20 回拡張を行った結果から、次の特徴があることが分かった。

- 1) 提案手法 1 の拡張子は 3 手法の中で最も長くなる傾向が強く、また、多様な拡張子が導出される。
- 2) 提案手法 2 の拡張子は 3 手法の中で最も短くなる傾向が強く、また、拡張子の多様性は少ない。

5. 関連研究

タキソノミーを用いた問合せ拡張に関する研究として文献[3, 4]がある。しかし、これらの方法では、拡張子はタキソノミーのノード情報のみで決定される。我々が検討している手法は、問合せに応じて動的に拡張子を決定する点でこれらとは異なる。文献[5]では Web 検索精度の向上のための Web ページの自動分類について検討している。しかし、問合せ拡張については考慮していない。文献[6]ではディレクトリ更新のための対話型問合せ学習システムを提案したが、その主な焦点は問合せの生成である。文献[7]では、Web ページに対する分類器作成において、正例とラベルのないサンプルデータから正例と負例を分類する SVM を導出する手法を提案している。これらの

表 1 . 実験に用いる問合せ

Query condition	Context category	Meaning
String	/Arts/Music/	Find pages about stringed instruments , stringed instrument music and musicians, string making, etc.
Salsa	/Arts/Performing Arts/	Find pages about Salsa dance including ballrooms, education, studios and instructors for Salsa dance, but they may also include information on other dances.
Oil	/Shopping/Health/	Find pages selling health oil products including beauty oil products, aromatherapy (essential oils), acne oil, etc.
Laser	/Health/Medicine/	Find pages about practitioners, clinics and resources on laser treatment in medical (e.g. , laser surgery).
Tent	/Recreation/Outdoors/	Find pages about tents in the context of outdoor recreation such as camping.
Apple	/Home/Cooking/	Find pages about apple cooking but not pages selling apple food products.
Oil	/Business/Mining and drilling/	Find pages about companies and products related to oil in the context of mining and drilling business.
Sun	/Science/Technology/	Find pages about the sun in the context of technologies and related applied sciences.
Metal	/Arts/Music/	Find pages about metal instruments, heavy metal music, heavy metal sound files, etc., but not pages selling metal instruments.
Nepal	/Recreation/Travel/	Find pages including travel information on Nepal.

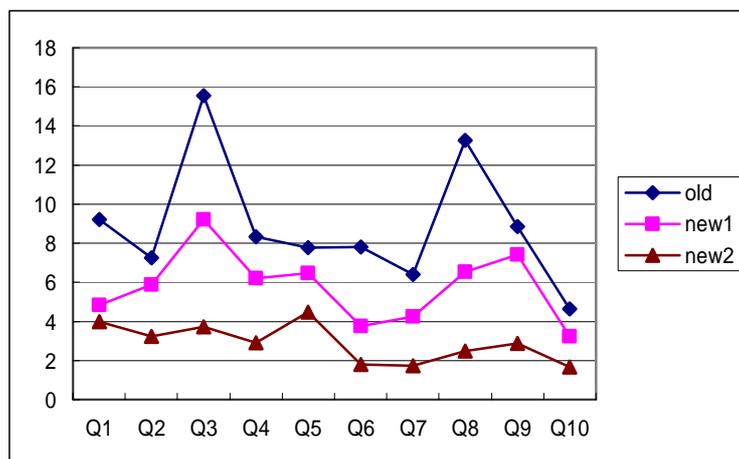


図 2 . 問合せ拡張子導出時間

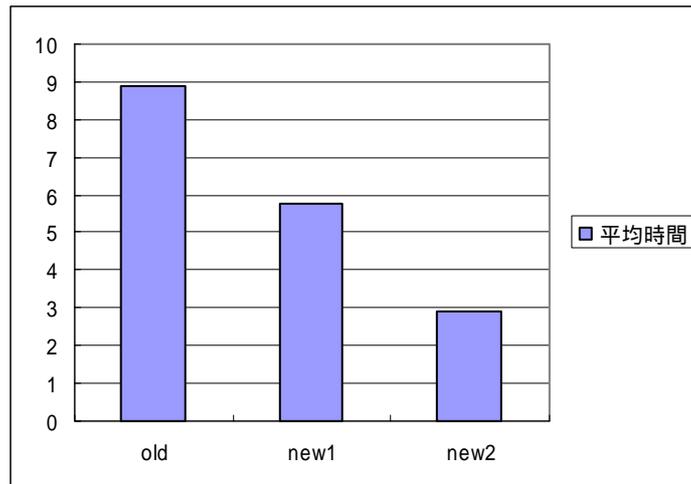


図3．平均時間

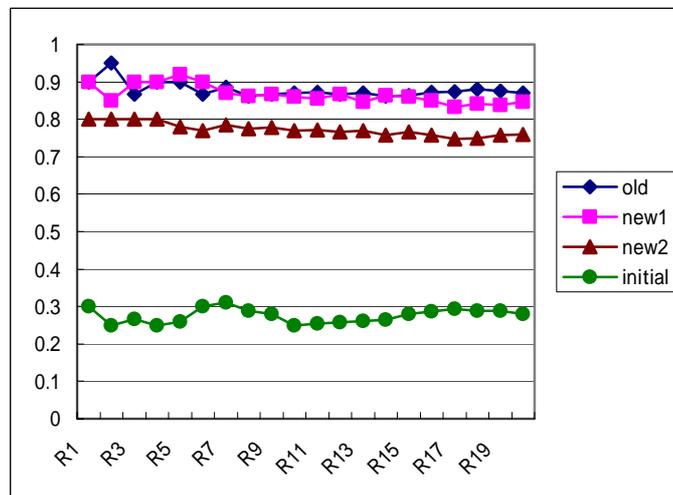


図4．適合率

手法と本提案手法の融合は興味ある今後の課題である．分類器作成における正例と負例の数の影響については，文献[8]などで検討されている．

6. まとめと今後課題

我々の研究グループでは，ディレクトリ型検索エンジンと対象検索エンジンを併用することで検索精度を向上させる，Web 統合検索方式の研究を行ってきた．本論文では，ディレクトリ型検索エンジンより動的に獲得する情報の量を削減する改良手法を2種類提案し，実際のWeb検索サイトを用いた実験により改良手法の有効性を評価した．この結果，全体の間合せ処理時間の短縮と，従来手法とほぼ同程度の検索精度の両立が可能なことを確認した．

今後の研究課題としては，以下の点があげられる．

- 1) より多様な状況での実験評価．特に，コンテキストカテゴリがタキシノミーの深い位置にある場合の評価が必要である．
- 2) 他の負例の取得方法についても，検討の余地がある．例えば，3節で述べたような点について検討する必要がある．

謝辞

本研究の一部は，科学技術研究費補助金基盤研究(B) (15300027)ならびに特定領域研究(2) (15017207)の助成による．

参考文献

- [1] S. M. Pahlevi and H. Kitagawa, "A Web Search Method Integrating Taxonomy-based and Crawler-based Search Engines", 情報処理学会論文誌：データベース, Vol. 43, No. SIG9, pp.15-27, 2002 .
- [2] S. M. Pahlevi and H. Kitagawa, "Building a Taxonomy-based Integrated Web Search Service", 第 14 回データ工学ワークショップ(DEWS2003), 2003.
- [3] E. J. Glover, G. W. Flack, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles, and D. Pennock, "Improving Category Specific Web Search by Learning Query Modifications", Symposium on Applications and the Internet (SAIT), pp.23-31, 2001.
- [4] S. Oyama, T. Kokubo, T. Ishida, T. Yamada, and Y. Kitamura, "Keyword splices: A New Method for Building Domain-specific Web Search Engines". IJCAI, 2002.
- [5] C. Chekuri and M. H. Goldwasser "Web Search Using Automatic Classification", Sixth International WWW Conference (WWW6), 1997.
- [6] W. W. Cohen and Y. Singer "Learning to Query the Web", Workshop Internet-based Information Systems 13th Nat. Conf. Artificial Intelligence, PP.16-25, 1996.
- [7] H. Yu, J. Han and K. C-C. Chang, "PEBL: Positive Example-Based Learning for Web Page Classification Using SVM", ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases (KDD02), ACM Press, pp.239-248, 2002.
- [8] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies", International Workshop on Learning from Imbalanced Data Sets, 2000.