入札情報検索システムのためのWebマイニング技術を用いた情報フィルタリング技術の開発

Development of Information Filtering Technology Based on Web Mining Technology for a Bidding Information Search System

小俣尚泰＊　関根聡一＊
Naoyasu Omata,　Soichi Sekine

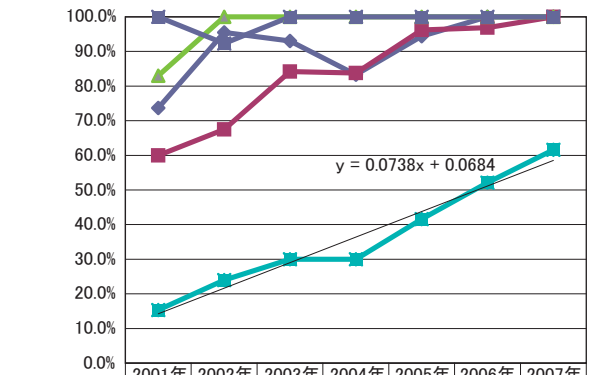
我が国では入札契約適正化法の施行以来、入札情報のWeb上での公開が進んでいる。これにより、各事業者がより早く、より広く情報を得る機会が増したが、発注機関ごとに公開方法が異なるため、利便性が高いとはいえない。そこで本研究では、Webを通じて公開された入札情報を横断検索できる入札情報特化型のWeb検索システムを構築した。本システムは発注機関ごとに異なる用語や文書形式の違いを吸収し、かつ収集しなければならない情報を判断する情報フィルタリング機構を特長とする。

Bid announcements on the Web have become popular in Japan since the "Act for Promoting Proper Tendering and Contracting for Public Works". But, the method for announcing bid information is not the same with every purchaser. Therefore, to effectively obtain bid information on the web is requires search costs. This paper reports on the development of a Bidding Information Search System and bid information filtering technology based on web mining technology. The feature of the filtering technology is the resolution of differences of a terminology and document types, and the extraction of information necessary from purchaser's web sites.

1. はじめに

1.1 インターネット上の入札情報の増加

入札契約適正化法の施行以来、発注見通・入札公告・落札結果などの入札情報のインターネット上での公開が進んでいる。図1に示すように、国土交通省の調査によると、2007年時点では、入札情報のインターネット公開が国・都道府県機関では100%、市区町村機関では61.6%となっている。これは年々増大する傾向にあり、インターネット上に確かに入札情報が存在している状況といえる。



	2001年	2002年	2003年	2004年	2005年	2006年	2007年
国	73.7%	95.5%	93.0%	83.3%	94.4%	100.0%	100.0%
特殊法人	60.0%	67.5%	84.2%	83.8%	96.2%	96.9%	100.0%
都道府県	83.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
指定都市	100.0%	92.3%	100.0%	100.0%	100.0%	100.0%	100.0%
市区町村	15.3%	24.0%	30.0%	30.0%	41.6%	52.1%	61.6%

図1　入札情報のインターネット公表の浸透状況

1.2 発注機関の情報インフラ投資上の問題

発注機関によるインターネット上での入札情報公開手段の標準化の流れとして、入札情報サービス（以下、統合PPI）への統合が推進されている[1)、2)]。全国の地方自治体を含む全ての入札情報の横断的検索サービスの提供には、この統合PPIへ情報が統合されるのが理想である。

しかしながら、その統合に必要となる参加発注機関側でのシステム導入は、特に市区町村などの小規模な機関において、費用対効果を期待できず導入が進まないという指摘がなされている[3)]。情報インフラ投資面での問題から、より安価なコンテンツ管理システムによる情報公開を選択する地方自治体が少なくなく、これが統合PPIへの統合を阻害する要因であると考えられる。完全な入札制度の電子化を目指すには、発注機関におけるシステム導入コストの低減を考慮しなければならない。

1.3 開発の経緯

当社を含む官公庁・地方自治体などを顧客とする公共向けの事業者（以下、受注者）にとって、発注に関わる入札情報を可能な限り広範囲かつ迅速に入手することが重要である。そのため、各種メディアからこまめに情報をチェックし、あるいは実際に顧客と折衝することにより情報を得ることが不可欠であるが、多大なコストを割くことになる。そこで、情報入手に関わるコストを低減することによる当社内での受注機会の向上、並びに地方自治体から公開される入札情報検索サービスの事業化を想定して、「早く」「正確に」「漏れなく」営業品目としてあげられている案件のみをチェックできる仕組みを実現する試みとして、入札情報検索システムの開発を進めている[4)]。

＊　技術開発本部 情報技術グループ

2．入札情報検索システムの要件

2.1 検索システムの基礎事項の整理

a)検索エンジン

一般的なWeb検索エンジンは、検索に対して早く結果を返す性能、すなわち応答性能を高めることを目的として、大きく分けて３つの要素からなる。まず、利用者からの検索要求に応えて検索を行うコンピュータ、すなわち「検索サーバ」と呼ばれるサーバを配置する。次に、インターネットから情報を集めて整理するコンピュータがあり、「検索バックエンド」と呼ぶ。最後に、それら二つの間で利用されるデータベースとなる「インデックス」が存在する。ここで、インデックスは検索対象となるべきインターネット上の原データ集合に対する写像であり、Webページ内に存在する情報を抽出し、検索のためにあらかじめ構造化したデータである。検索サーバはユーザからの検索要求に対して、インデックス内を検索する。すなわち、検索サーバはインデックスを使用するサーバである。検索バックエンドは、後述のクローリングの処理を含むインデックスを作成するまでの処理を行う。

b)Webクローラ

Webクローラとは、インターネット上のWebサイトからハイパーリンクを探索しながらWebページを収集する機構をいう。図２に示すように、探索の起点を定めて、次に収集するべきWebページを順次取得して処理が進んでいく。クローラが集めたWebページは一時的に「リポジトリ」と呼ばれる領域に保管する。この処理をクローリングという。次にWebクローラが集めてきたWebページからインデックスを作成する。インデックスの作成では、Webクローラの動作ログ解析、単語の解析

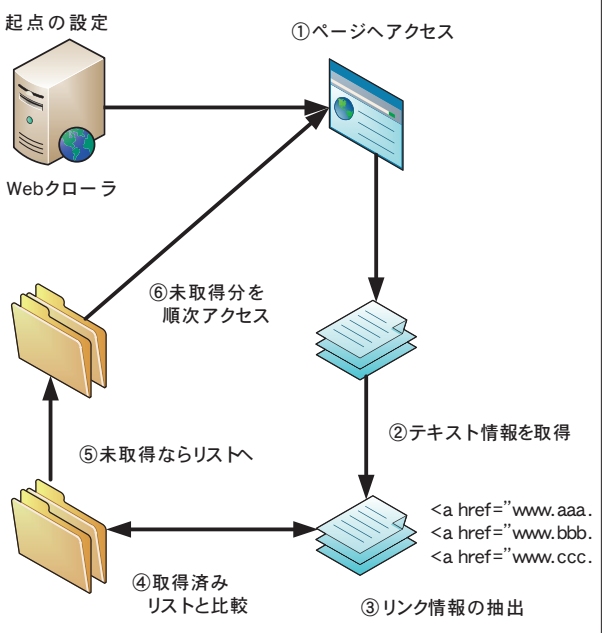


図２　Webクローラの動作

処理、ハイパーリンク構造解析を通じて、Webページからユーザの検索要求と照合するためのデータを抽出する。以上の動作原理により、検索エンジンで提供される検索結果は、Webクローラがある時点で集めて来たインターネット上のデータに対しての検索を行うことになる。原データにより近づいた検索対象に対して検索を行えるようになるのは再度クローリングが行われた後となる。

2.2 入札情報に特化した検索システムの要件

本研究では、Webサーバを通じて公開されるPDF形式などの文書を収集・分析し、入札情報に特化したWeb検索エンジンの構築を目指す。そこで筆者らは、入札情報の検索システムとしての要件を以下の通りに規定した。

・情報の鮮度を保つこと
・情報を正確に探し出せること
・情報の網羅度を高くすること
・利便性の高いシステムとすること

以下では、これらのシステム要件について、一般的なWeb検索エンジンの振舞いとの比較から課題解決策を述べる。

a)情報の鮮度を保つこと

情報の鮮度とは、情報の更新頻度であり、インデックスの更新遅延時間を指す。例えば、入札公告は入札実施の１週間程度前に掲載され、入札が終了すると削除され入札結果の文書として更新をされる。ここで、この入札公告を一般的な検索エンジンで、検索結果として検索することを考える。検索バックエンドがコンテンツのダウンロードからインデックスの作成までを行う時間は一定時間を要する。そのため、先述のWebクローラの動作原理により、ある時刻において検索要求を検索サーバが受けたとき、インターネット上の原データと検索エンジンが実際に検索を行うインデックスには必ず更新遅延が生じている。ここで、あるWebサイトW内のWebページが$w_i \in W(i=1,2,\ldots,n)$Webクローラによりダウンロードされインデックスの作成が完了するまでの時間を$t(w_i)$とおく。また、一般的なWebクローラでは、大規模な収集を効率よく行うため、w_iに対して図２に示される収集を再度実行するまでの間隔を調節している。この再収集が行われるまでの遅延時間を$\alpha(w_i)$とおく。以上より、あるWebページがインデックスに登録されるまでに要する時間$T(w_i)$は次のように定式化できる。

$$T(w_i) = t(w_i) + \alpha(w_i)$$

一般的な検索エンジンでは、ユーザに提供される$t(w_i)$および$\alpha(w_i)$には、一定時間以内に完了するという制約および保証はなされない。そのため、あるWebページw_iに対して$T(w_i)$は１日であったり、１週間以上であったりと変動をする。そのため、一般のWeb検索エンジンでのWに対するインデックスW'にて入札情報のチェックを行うと、ある一定期間で結果として出るべ

き入札公告が検索結果に出てこないことや、入札情報はなかったということが保証できないケースが生ずることが予想される。この問題の解決のため、$T(w_i)$を数日程度以内という保証を採り更新遅延時間を少なくすることや、あるWebサイトWに対しては、更新遅延$T(w_i)$は既知の上で、一定期間内の「情報はなかった」という結果を検索エンジンが提供できるような検索バックエンド処理を構築することにより情報の鮮度を高める必要がある。

b)情報の正確性を保証すること

入札情報という限定的な用途であるがゆえに、検索エンジンによる結果の正確性についても問題が生ずる。一般の検索エンジンはさまざまなニーズに対応すべく汎用的につくられているため、入札情報の調査業務に適用するには、検索時に設定できる項目が少ない。そのため得られる検索結果は、再現率は高いが適合率が低く、不必要な結果が多数混ざることとなる。例として、当社内でのアンケート結果(図3、図4)によると、工事名称と内容説明に対する全文検索や、予定価格の上下限範囲を指定しての検索などに強いニーズがある。また、入札結果に対しては、落札者や落札金額を指定して検索することにより競合他社の動向を分析したいというニーズも強い。したがって、一般のWeb検索エンジンよりも入札情報に特化した検索項目を備える必要がある。

c)情報の網羅度を高くすること

情報の網羅度を考えると、検索するデータ母集団は、受注者側が営業範囲としている発注機関を網羅することが必要である。入手しなければならない情報は、発注見通・入札公告・入札結果の3つの情報区分と、発注機関である。ここでは、対応する発注機関を増やせば増やすほど、システムが取扱うデータの規模が大きくなる。クローリングしたWeb文書の網羅度のみを考えれば、一般の検索エンジンでも実現は可能であるが、先述の鮮度・正確性が担保できない。また、独自に設計したWebクローラでは、不要なデータが検索のデータ母集団となってしまうという、データベース構築上の問題が生じ、検索性能に影響する。

d)利便性の高いシステムとすること

その他の要件としては、利便性の高いシステムであることが挙げられる。利便性についての具体的な要件を以下に述べる。

1)可用性が高いこと

可用性とは使いたいときに使えるかどうかという性質である。ここでは、特別なトレーニングを受けることなく誰でも簡単に利用できること、使用するにあたり特別な環境を要せずとも使用できることが求められる。したがって、簡易なユーザインターフェースを備えるWebアプリケーションが望ましいと考える。

2)可搬性があること

外出の多い営業職からは、モバイル機器からのアクセスに強いニーズがある。担当エリアの営業方針にしたがった情報を、出先でも利用できる環境が整うことはビジネスのスピードを上げるためにも欠かせない。

3)情報を自動的に知らせること

本研究は検索システムの実現を目指すものであるが、理想的にはユーザが欲する情報を自動配信することが望ましい。ここで、期中での営業方針に変更がなければ、同一ユーザの担当する営業品目や営業範囲は通常同じであると考えられる。そのため、ユーザごとに検索条件は固定される。したがって、指定された検索要求に対する適合率を高く保つことができれば、ユーザごとに適した情報をモバイル機器へ自動配信することが可能となる。

必ず欲しいと思う	◎
あれば良いと思う	△

条件項目	発注見通	入札公告	入札結果
公表日・公開日	◎	◎	◎
発注時期	◎		
工事名称	◎	◎	◎
工事場所			
工期			
工事概要・内容説明	◎	◎	◎
発注方式(単体・ＪＶ等)			
参加資格			
申請書交付日			
申請書交付場所			
予定価格	◎	◎	◎
入札執行日	△	△	△
設計図書			
落札者			
落札金額			
入札参加者			
入札参加者の応札額			
入札日			
契約日			

図3　入札情報検索システムに必要だと思う検索項目(工事予算担当)

条件項目	発注見通	入札公告	入札結果
公表日・公開日			
発注時期			
工事名称	◎	◎	◎
工事場所			
工期		◎	△
工事概要・内容説明	◎	◎	◎
発注方式(単体・ＪＶ等)			
参加資格		◎	
申請書交付日			
申請書交付場所			
予定価格	△	◎	◎
入札執行日			
設計図書		△	
落札者			◎
落札金額			◎
入札参加者			
入札参加者の応札額			
入札日		◎	
契約日			

図4　入札情報検索システムに必要だと思う検索項目(設計担当)

3. 入札情報検索システムの構成

3.1 概要

本章では、上述の用件を満たす入札情報検索システム（以下、本システム）の構成について述べる。図５に示すように、本システムでは処理内容により、①データ収集部 ②データ加工・解析部 ③データ認識部 の３段構成を採る。以下、その具体的構成方法について述べる。

3.2 各部の構成

a）データ収集部

ここでは、Webクローラを通じて発注機関のWebサイトから入札情報を入手する処理を行う。クローラの処理結果から入札情報のWebページとリンク構造のデータが得られる。

b）データ加工・解析部

ここでは、前項a）で得られたWebページとリンク構造のデータの解析を行う。具体的には発注機関と情報区分の組に対する特徴の分析を行い、入札情報の抽出に特化した識別器である入札情報フィルタを作成する。この特徴の分析を行う際には、得られているWebページが、どの発注機関であり、どの情報区分に属するかを示す事前情報を作成する必要がある。

c）データ認識部

ここでは、前項b）の手順により得られた情報フィルタを使用して、再度行われるクローリングの際に未知となるWebページに対して、情報フィルタリングを行う。このような構成を採ることで、クローラが収集をしてくる不必要な情報を排除し、必要な情報のみを検索DBに登録されるようにする。

3.3 本システム構成によるメリット

入札情報フィルタを搭載したWebクローラにより適切に情報の仕分けを行うことで、一般的なWeb検索エンジンで実現困難な情報の鮮度・正確性・網羅度の問題を解決することができる。また、Webクローラ方式により、対象となる入札情報はWebサーバに公開されたWeb文書から取得できるため、発注者側での情報インフラへの投資効率の面で良い効果が得られる。

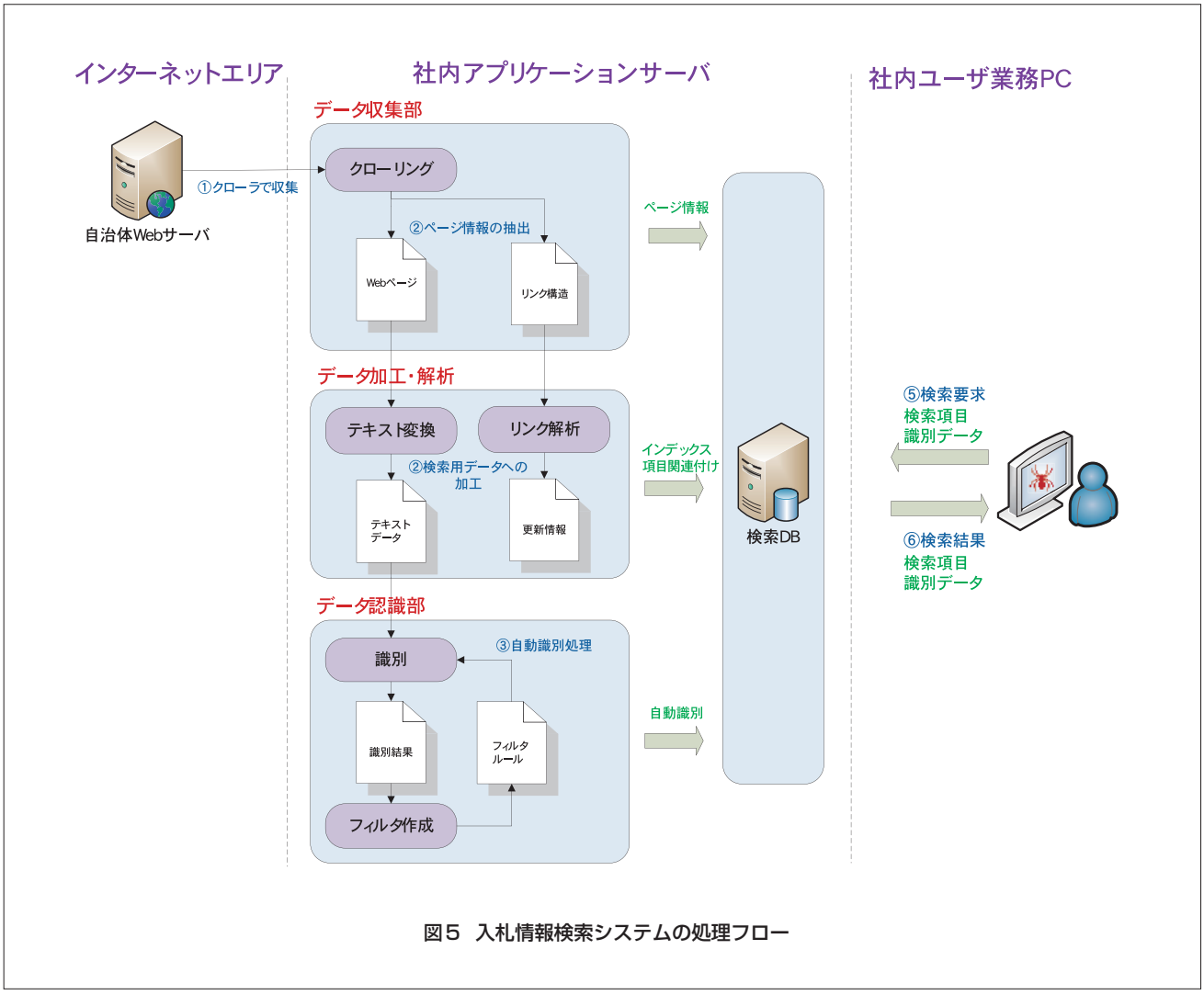


図５　入札情報検索システムの処理フロー

4. 入札情報フィルタの開発

4.1 Webクローラの特性上の問題点

a)動作特性上の問題点

Webクローラの収集動作の特性上、取得対象となったWebページが全て収集され、不要な情報が検索データベースへ登録されてしまう。そのため、入札情報検索システムを実現するためには、必要な情報のみを判断する機構を備えた独自のWebクローラが必要となる。

b)情報フィルタリングの導入による解決

情報フィルタリングとは、大量の情報の中から、ユーザにとって必要な情報を取出し、不要な情報を除外する処理を自動的に行う技術のことをいう。要・不要の2つに分けるほかに、情報に重要度や類似度などのメタ情報を加えて重み付けを行うものも含まれる。実用されている例としては、迷惑メールの除去フィルタ[6]などがある。図6に示すように、情報フィルタリングは未知の情報に対しての仕分け作業といえ、この仕分けのルールを特定することによってフィルタを作成することができる。

前項の要求を満たすため、この情報フィルタリング機構をWebクローラが備えることによって、情報の選別を自動的に行う入札情報フィルタを構成することができると考える。

4.2 Webマイニングによる情報フィルタリング

前節の入札情報フィルタの開発にあたっては、どのような情報フィルタリング手法を確立するかが入札情報検索システムを実現する上で重要となる。Webクローラにより収集をしたデータは膨大な量となり、効率良く必要な情報を抽出する処理が必要となる。近年では、Webシステムに関わる多種多様なデータに潜むパターンやルールの発見を目標にWebマイニングと呼ばれる研究が行われている[5]。本開発では入札情報では、Webマイニング技術を活用し、Webページから抽出した入札情報と判断するため特徴となるデータを処理することで入札情報フィルタの実現を目指す。Webマイニングは、その対象とするデータの特徴を踏まえて、以下の3領域に大きく分類されている。

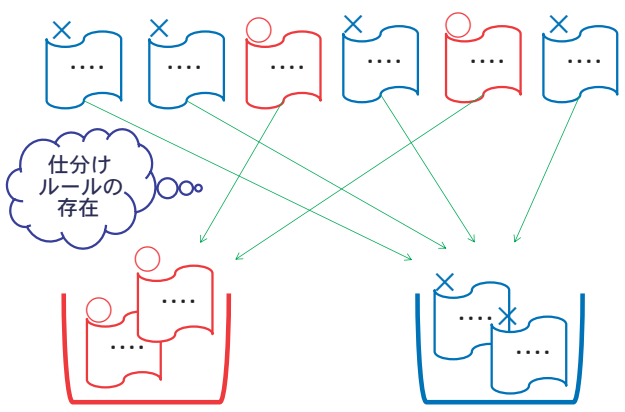


図6　Webページ分類のイメージ

a)Webコンテンツ・マイニング

Webコンテンツ・マイニングは、Webコンテンツ内の各種マルチメディアデータからルールを見つけ出す手法であり、Webページの特徴ベクトル生成を行う前処理を要する。そこで、非構造データとしてのテキストデータに対する自然言語処理技術を用いた解析や、HTMLやXMLによる半構造データのタグ記述を利用したキーワード抽出も行う技術である。内容をもって判断を行う技術として、入札情報フィルタの核となる。

b)Web構造マイニング

Web構造マイニングは、Webページ群を結ぶハイパーリンクからなるWebグラフ解析に基づく特徴の発見的手法であり、興味を同じくする利用者群を見出すコミュニティ形成や、Webページ群やコミュニティ内の代表ノードの選出に用いられる。図7のように、ある発注機関の入札情報は、リンク構造上どこに良く現れるか等を解析することに用いることができる。

c)Web利用マイニング

Web利用マイニングは、不特定多数のユーザによるWebアクセスのログや、ブックマークなどに記録された行動履歴から、利用者のアクセスパターンやブラウジング目的などを見出すものである。

4.3 フィルタの作成単位

本稿で開発を行う入札情報フィルタは、教師あり学習機械であり、学習を行うための教師データを必要とする。教師データを作成する上で、前節で述べた仕分け作業を行う単位を次の2種類とした。以下にそれぞれの作成方法について述べる。

a)タイプ1：情報区分による仕分け単位

ここでは「発注見通」「入札公告」「入札結果」を情報区分としている。Webクローラにより集められて来たWeb文書に対して、この情報区分を付与し、仕分けを行う。

例えば、このWeb文書は「発注見通である」「発注見通ではない」という2値のメタデータが付与されることになる。この仕分け単位で見た場合、Webクローラに設定された発注機関内の範囲において、どのような文書形式で情報が掲載されているかの全体の傾向を掴むことができる。この仕分けは、3.2節で述べたデータ加工・解析部によって自動的に行う。

b)タイプ2：発注機関×情報区分による仕分け単位

ここでは、Web文書に対し、どの発注機関のものであるかの情報を付与する。先述の情報区分と同様に、このWeb文書は「○○発注機関より発行された文書である」「○○発注機関より発行された文書ではない」というメタデータを付与する。ここでは、先述の情報区分と組合わせて、あるWeb文書は「○○発注機関より発行された文書であり、発注見通しである」という表現となり、4値での判定となる。この方法は、文書作成業務などのマニュアル化およびWebサイト作成上の仕様は発注機関

ごとに存在するため、Web文書の特徴は、発注機関サイト内で同定するべきであるという仮定に基づいている。この情報区分の付与は、手動による作業にて行う必要がある。

4.4 入札情報フィルタ

前節の教師データを基に、表1に示す6種類の入札情報フィルタの設計と作成を行った。あるWebページに対して、どの特徴をもって入札情報と判断するかの判断基準がそれぞれのフィルタで異なっている。それぞれのフィルタは作成時にLeave-one-out法[7]により精度の検証を行う。

a)テキストフィルタ

あるWebページ内のテキストデータにおける形態素列をベクトルデータ化したものから特徴ルールを抽出する。分類器にはベイジアンフィルタ[8]を採用した。テキストフィルタは、仕分け単位ごとに2種類作成し、テキストフィルタ(タイプ1)とテキストフィルタ(タイプ2)とした。このフィルタでは、例えば入札結果を判定する場合、「契約」「落札金額」などの入札結果を表現するのによく用いられる語句が含まれているWebページを入札結果と判定するという動作をする。

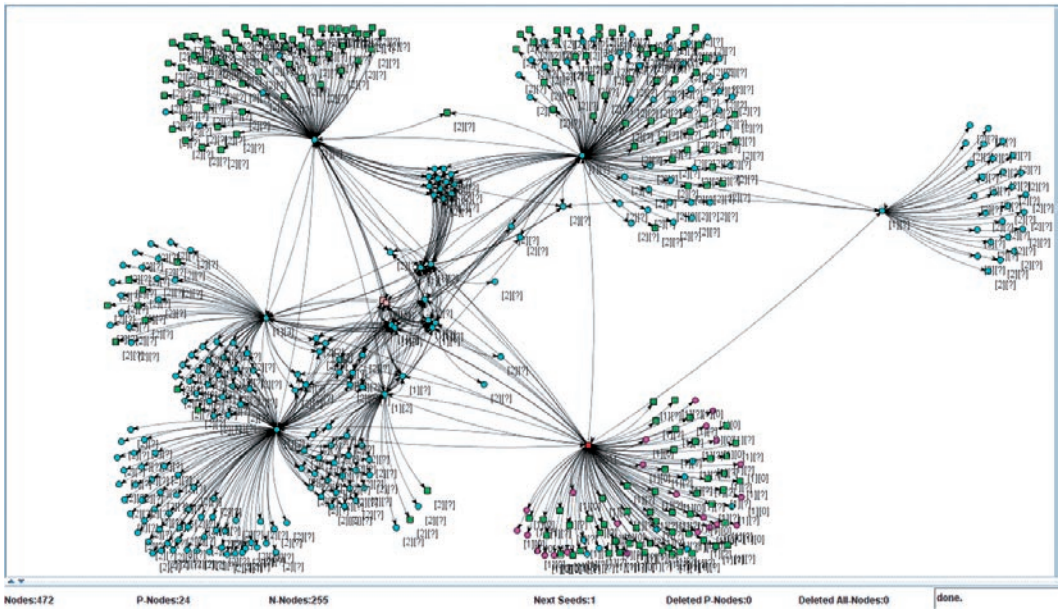


図7 Web構造マイニングの例

表1 入札情報フィルタと教師データの対応

フィルタ名称	教師データの仕分け単位
テキストフィルタ(タイプ1)	タイプ1
テキストフィルタ(タイプ2)	タイプ2
URLフィルタ	
リンク構造フィルタ	
ファイルタイプフィルタ	
複合フィルタ	

図8 入札情報フィルタ作成のための情報の仕分け単位

b)URLフィルタ

あるWebページを表す一意のURL内において、一定の意味をなすトークン列をベクトルデータ化したものから特徴ルールを抽出する。仕分け単位はタイプ２のみ使用した。分類器にはベイジアンフィルタ[8]を採用した。このフィルタでは、例えば、URLに「nyuusatsu」など入札情報を特徴付けるようなトークンが含まれる場合、入札情報と判定するという動作をする。

c)リンク構造フィルタ

あるWebページにおいて、リンクの「入次数」「出次数」「Webクローラの探索起点からの距離」の３次元データから特徴ルールを抽出する。仕分け単位はタイプ２のみ使用した。分類器にはクラス重心からのユークリッド距離を用いた。このフィルタでは、例えば、ある発注機関において、入札情報を掲載したWebページに存在するリンクの数が他のWebページに比べて少なかった場合、それらを元に判定するという動作をする。

d)ファイルタイプフィルタ

あるWebページにおけるHTMLやPDFなどのファイルの種類別の割合を特徴とする。仕分け単位はタイプ２のみ使用した。分類器には、しきい値方式を用い、仕分け内において半数を超えて入札情報に用いられるファイルタイプは入札情報として判定する。

e)複合フィルタ(多数決方式)

複合フィルタとは、タイプ２の仕分け単位内において、a)のテキストフィルタ(タイプ２)とb)～d)のフィルタの４つを用いて、それぞれ後述するF値を重みとして、フィルタを結合したものである。

4.5 精度予測実験

前節の方法による入札情報フィルタを作成し、精度予測の実験を行った。日々更新される未知のWebページへ追随するための精度検証を行う。テキストフィルタ(タイプ１)は学習データが多いため再代入法により検証を行い、その他のフィルタはLeave-one-out法により検証を行った。

a)Webページの収集

次に示す対象発注機関のWebサイトごとに入札情報へ到達できる探索起点を定め、Webページを収集した。探索深さは起点から２とし、この条件で到達できるWebページを実験対象とする。

対象発注機関：水道事業体353機関

収集したWebページ総数：189,453

b)入札情報への仕分け作業

前節で述べたフィルタを作成するための教師データとして、情報区分の情報を人手により付与する。本実験では、情報区分は「発注見通」「入札公告」「入札結果」とし、あるWebページが、いずれに属するかを判定した。タイプ２の仕分け単位の最大数は、単純には次式で求められる。

[タイプ２総数]＝[発注機関数]×[情報区分の数]

しかしながら、ある発注機関では、「発注見通」「入札結果」はWebで公開しているが「入札公告」は公開していないなどの対応があるため、今回はタイプ２の仕分け単位の総数は868となった。

c)入札情報フィルタの作成

前項b)にて、作成した教師データをもとに、入札情報フィルタの学習を行う。作成されるフィルタは次の６つであり、最も精度が高いものを、その仕分け単位で使用するフィルタとして採用する。

・テキストフィルタ（タイプ１）
・テキストフィルタ（タイプ２）
・URLフィルタ
・リンク構造フィルタ
・ファイルタイプフィルタ
・複合フィルタ

ただし、フィルタが適用できない状況を防ぐため、次の条件では、強制的にテキストフィルタ(タイプ１)を採用する。

・テキストフィルタ（タイプ１）以外のフィルタが全てF値0.5以上を確保できない場合。
・タイプ２の仕分け単位内のデータが少なすぎて、テキストフィルタ（タイプ１）以外のフィルタの作成が行えない場合。

4.6 実験結果と考察

実験結果を図９～図12に示す。

a)フィルタの選択状況

図９では、Multipleが複合フィルタ、URLがURLフィルタ、Fileがファイルタイプフィルタ、Textがテキストフィルタ(タイプ２)、Linkがリンク構造フィルタ、Text-V1がテキストフィルタ(タイプ１)の採用数と割合を示す。それぞれの仕分け単位内において最も精度の良いフィルタを選択した結果である。ここでは、仕分け単位で見ると、タイプ１を使用するものが47.12%で、タイプ２を使用するもの52.88%となっている。また、タイプ２を使用するものの内訳を見てみるとリンク構造

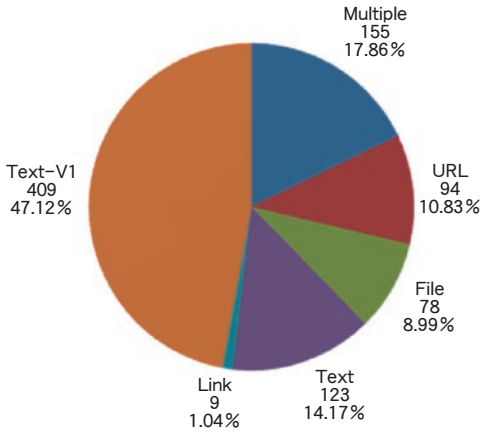


図9　各種の入札情報フィルタの選択割合

フィルタが適用されているところは少数ではあるが、5つのフィルタがそれぞれ採用されていることから、URLを見れば入札情報であると判定できる発注機関であったり、ファイルタイプを見れば入札情報であると判定できる発注機関であったりと、その仕分け単位内で適したフィルタが異なっているということがわかる。これにより、入札情報を見つけだすためには、複数の特徴を検討する必要があることがわかった。

b)フィルタの精度

タイプ2の仕分け単位ごとに採用されたフィルタの精度をプロットしたものを図10に示す。タイプ2仕分け単位を使用するフィルタの総数は、前項a)より459ほどある。性能が良いフィルタを得るためにはそれぞれの点がより右上に近づく必要がある。

縦軸は再現率で、「入札情報」を正しく「入札情報である」と判定できる性能である。この評価値が高いフィルタは収集してきた入札情報を漏れなく検索データベースへ送ることができる。入札情報であるWebページに対してフィルタが正しく入札情報であると判定した数を*TP*、入札情報であるWebページの総数を*P*とおいたときの再現率*TPF*を次式に示す。

$$TPF = \frac{TP}{P}$$

横軸は適合率で、「入札情報ではない情報」を正しく「入札情報ではない」と判定できる性能である。この評価尺度が高いフィルタは、収集してきた不要な情報が検索データベースに登録されることを防ぐ性能が高くなる。入札情報ではないWebページに対してフィルタが正しく入札情報ではないと判定した数を*TN*、入札情報ではないWebページの総数を*N*とおいたときの適合率*TNF*を次式に示す。

$$TNF = \frac{TN}{N}$$

フィルタの精度をF値で表し、度数分布として示したものを図11に示す。ここで、再現率を*TPF*、適合率を*TNF*とおくと、F値は次式で表される式であり、フィルタの精度を1つの性能尺度で見ることができる。

$$[\text{F値}] = \frac{2 \times TPF \times TNF}{TPF + TNF}$$

F値=1.0を達成しているものは179であり、総数の39%である。この結果は、タイプ2の仕分け単位で学習を行うことができるフィルタを作成しようとした場合、F値=1.0のフィルタの成功率は約39%ということを示す。また、仕分け単位の総数の割合で見ると、総数の80%がF値=0.83以上、総数の90%ではF値=0.74以上を確保できるという結果となった。F値の低い領域のフィルタが少なからず存在しており、今後はこれらの領域となったフィルタの改善が必要となる結果であった。

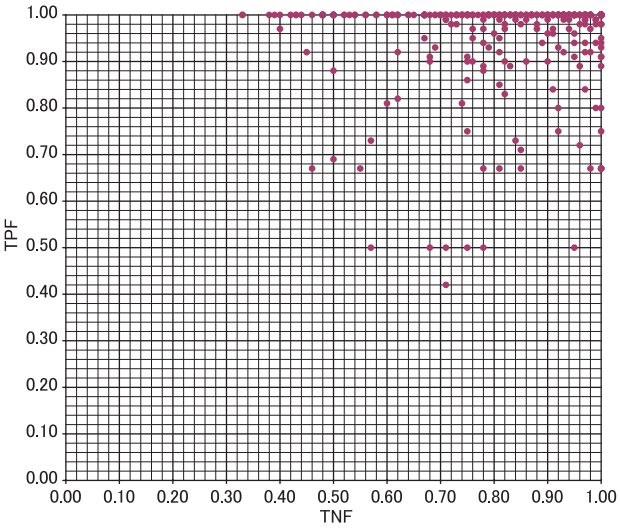


図10　入札情報フィルタの精度の分布状況

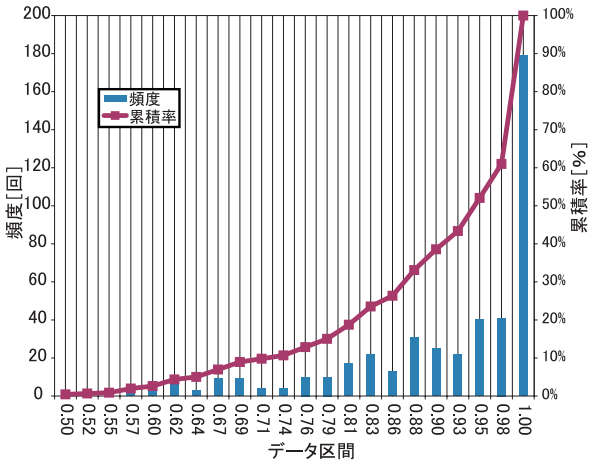


図11　入札情報フィルタ精度の度数分布（F値）

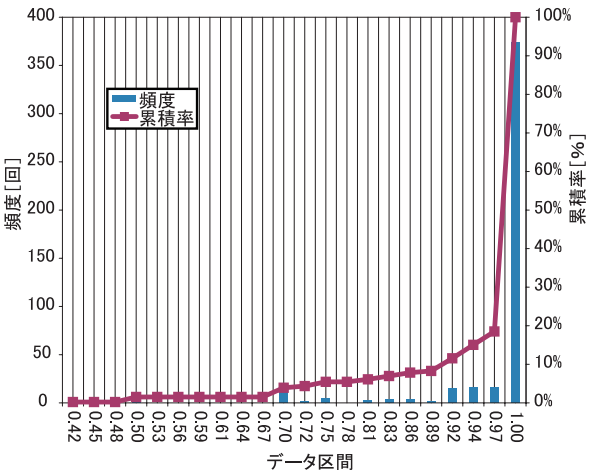


図12　入札情報フィルタ精度の度数分布（再現率）

図12は再現率を度数分布として表したものである。［再現率］=1.0を達成しているフィルタは374あり、総数の81%である。また、総数の90%が［再現率］=0.92以上を確保できている。この結果より本稿における入札情報フィルタは、多数の発注機関において比較的高い再現率を示すことがわかった。これは、入札情報を正しく入札情報であると判定する性能が良好であることを示す。

図13は適合率を度数分布として表したものである。［適合率］=1.0を達成したフィルタは190あり、総数の41%である。また、総数の90%が［適合率］=0.62以上を確保できている。適合率が低い領域にもフィルタが多数存在していることがわかり、この適合率の性能がF値を下げる要因となっていることがわかる。

4.7 実験まとめ

本章では、Webクローラが集めてきたWebページの中から入札情報のみを選別する入札情報フィルタの作成および性能実験を行った。実験結果から次のことを達成した。

a)自動選択情報フィルタ機能の開発

ある発注機関サイト内における学習データを分析し、そのサイト内で適する情報フィルタを、6種類ある候補の中から適するものを自動的に選択する機構を開発した。これにより、各発注機関ごとにどのような入札情報フィルタを構成すれば良いかを判断するためのフィルタ作成に要する分析作業を省力化することができる。

b)ユーザに検証値を提示することが可能

本稿で開発した入札情報フィルタは、全てのフィルタに対し、作成時の検証処理において、F値をあらかじめ算出することができる。そのため、検索システムを導入する上で、その性能尺度をユーザに提示することが可能になった。

c)現状の入札情報フィルタの特性

再現率では、ほとんどのフィルタが高い精度を持つため、欲しい情報の再現性は高いといえる。すなわち、入札情報検知漏れが少ないといえる。しかしながら、適合率が性能劣化の要因となっており、不要な情報が検索結果に混在する。これは、キーワード検索などで絞り込むことにより回避することができる。

5. Web検索システムとしての実現

4章で述べた入札情報フィルタを組込んだWebクローラにより登録を行ったデータベースに対して、検索機能を提供するユーザインターフェースの開発を行った。本システムでは、図14に示すように検索項目として使用できる項目は次の通りである。

・情報区分　（発注見通｜入札公告｜入札結果）
・キーワード　（任意の文字列）
・都道府県
・発注機関　（水道事業体353機関）
・ファイルタイプ
　（HTML|PDF|EXCEL|WORD|その他）※）
・更新日　（Webクローラが更新を検知した日）
・フィルタ条件　（前章のフィルタの適用）

入札情報の検索システムとしては、検索項目がまだまだ少ない状況ではあるが、ページ内の情報をより細かく抽出することで、発注金額や工種、入札執行日といった項目に対応していきたい。

図15は検索結果の例である。フィルタを適用しない場合は、Webクローラが集めてきた情報がそのまま検索される。フィルタをかけた場合は、前章の入札情報フィルタがかかり、不要な情報を検索結果から排除しようとする。前章に示す通り、適合率に課題があり、不要な情報が混在するという事象が発生するため、今後の開発により改良を重ねていきたい。

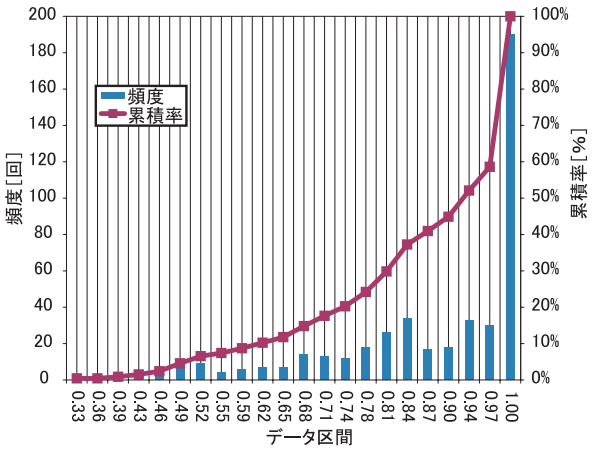


図13　入札情報フィルタ精度の度数分布（適合率）

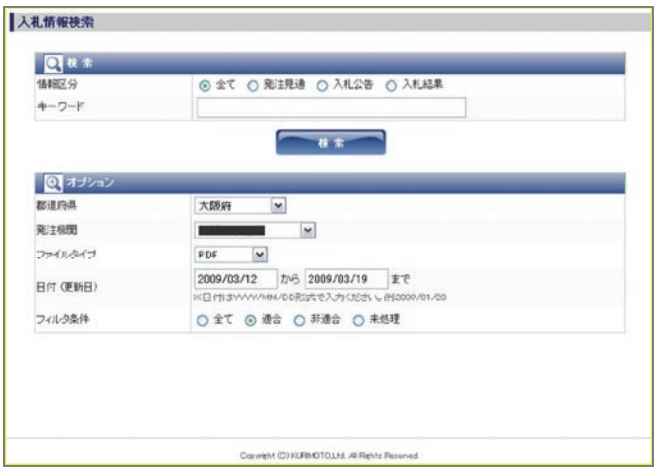


図14　入札情報検索システム（検索画面）

※）記載された製品は、各社の商標または登録商標です。

図15 入札情報検索システム(検索結果画面)

6. まとめと今後の課題

本稿では、入札情報の公開から取得までの流れを全体で効率化することを主眼に置き、Webクローラ方式による入札情報に特化したWeb検索エンジンのシステム構成を提言し、プロトタイプシステムを設計・開発した。また、そのメリット・デメリットに言及し、本システム構成は、発注機関・受注者双方にとって、情報インフラに対する投資効率を高める可能性があることを示した。また、そのWeb検索エンジンへ登録を行う際には、入札情報のみとなるよう、入札情報フィルタの開発を行った。

今後の課題としては、次のような事項が挙げられる。

a)情報フィルタの精度

情報フィルタの精度が全体の性能に大きく影響するため、その精度の維持は最重要の課題となる。そのフィルタは再現率に課題が残るものであり、精度向上が必要である。フィルタ作成時における特徴データをより詳細にするなどの施策が考えられる。

b)維持コスト

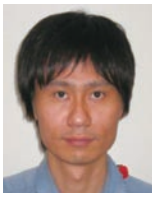
情報フィルタの精度を維持するためのもうひとつの要因として、3章で述べたデータ分析・加工部で必要とする事前情報の充実がある。この情報の作成は人手を要するため、本システムを維持するためのコスト要因となる。Webデータの量は多いため、少量の作業で高い精度を作成できる情報フィルタの構成法または事前情報の作成方法が必要となる。

参考文献

1)(財)日本建設情報総合センター:入札情報サービス、http://www.i-ppi.jp
2)国土交通省:入札情報サービス(ＰＰＩ)の移行について、http://www.mlit.go.jp/kisha/kisha07/13/130628.html
3)(財)日本建設情報総合センター研究助成事業:中小都市における電子入札制度の効果報告書、http://www.jacic.or.jp/kenkyu/5/5-2-2.pdf(2003)
4)小俣尚泰、関根聡一:Webマイニング技術による入札情報検索システムの開発、第34回情報利用技術シンポジウム講演予稿集、土木学会(2009)
5)坂本比呂志、有村博紀:ウェブ・マイニング、人工知能学会誌、Vol.16、No.2(2001)、pp.233-238
6)安藤一憲:フィルタリング、情報処理学会誌、Vol.46、No.7(2005)、pp.758-761
7) Keinosuke Fukunaga: Introduction to statistical pattern recognition, Academic Press (1990)
8)Paul Graham: Better Bayesian Filtering, http://www.paulgraham.com/better.html(2003)

執筆者

小俣尚泰
Naoyasu Omata
2006年入社
情報通信に関わる研究開発に従事



関根聡一
Soichi Sekine
1993年入社
情報通信に関わる研究開発に従事