Webクローラ方式による入札情報検索システムのための情報フィルタリング

小俣　尚泰†　　関根　聡一†

株式会社栗本鐵工所†

1. はじめに

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．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・都道府県では100%，市区町村では61.6%の機関で実施されている．実施機関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．そこで筆者らは，官公庁・地方自治体などを顧客とする公共向けの事業者に対して，発注に係わる入札情報を広範囲かつ迅速に入手することを支援する入札情報検索システム（以下，本システム）の開発を進めている．本システムは図1に示すように一般的なWebクローラ方式によるWeb検索エンジンの構成を採っている．本システムで問題となるのはWebクローラが収集したデータから利用者にとって必要な入札情報のみを抽出する入札情報フィルタの実現である．そこで筆者らは，内容テキスト・URL・ファイルタイプ・リンク構造の4つの評価観点を用いてWebサイト特性に応じて評価観点を選択的に変更することができる入札情報フィルタを開発し評価実験を行った[1]．良好な識別精度は得られたものの2クラス分類器である本フィルタを構成するためには多量の訓練用サンプルを準備しなければならないという問題がある．日々変化するWebサイトに追随するために，特に負例ラベルを付与するコストが大きい．そこで本稿では，準備コストの削減を目的として，正例のみのデータを使用する類似度によるフィルタリングを検討する．

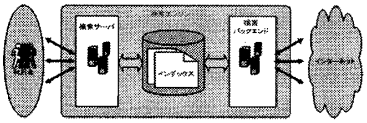


図1　Web検索エンジンの構成

2. 入札情報検索システムが持つ課題と解決策

本システムの目的は，図2に示すように自治体Webサイトにおいて目的領域内の入札情報を抽出することである．しかしながら，Webクローラはリンクを単純に辿る動作をするため，領域外の情報も多く集めてしまう．

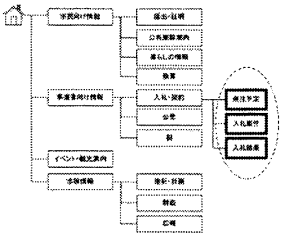


図2　自治体のWebサイトマップと入札情報

ここで，目的領域内かどうかのラベルをWeb文書に付与していくことを考える．2クラス分類器と類似度による情報フィルタリングでは表1に示す違いがある．2クラス分類器では，収集結果から抽出した訓練用サンプル全てに対して正負のラベルを付与する必要がある．精度良く識別ができた場合は，負例と判別された文書を削除する根拠を得ることができる．ただし，100%の識別精度が得られない場合は，誤識別された正例が検索インデックスに登録されず，インデックスの再現率が低下するリスクが顕在化する．一方，類似度による方法では，学習には負例ラベルが不要である．しかし，識別結果を根拠として負例ラベルが付与されるべきサンプルを除去するためには，正負の分布状況の情報を知る必要がある．本稿では以上のことを踏まえて，類似度による方法を検討する．具体的には，Webクローラが収集して来た範囲の文書は全てインデックスに登録し，正例クラスへの類似度をあらかじめ算出・保持することによって，正例クラス内文書であると推察される文書を検索結果の上位に表示するためのフィルタの用途を考える．

表1　識別方式の比較

識別方式	正例	負例	識別結果による負例の除去	再現率低下のリスク
2クラス分類	要	要	可能	ある
類似度	要	不要	事前情報が必要	ない

3. 類似度による入札情報フィルタリング

Web文書の類似度を算出するためにはベクトル空間への写像が必要となる．文書の評価観点は，内容・URL・ファイルタイプ・リンク構造の4つを使用する．表2に各評価観点のベクトル基底の選択，学習方法，類似度を示す．以降に具体的なモデルの定義を述べる．

表2　評価観点別のフィルタ作成方法

評価観点	対象データ	学習方法	類似度
内容	文字列	重心	内積
URL	文字列	重心	内積
ファイルタイプ	カテゴリカルデータ	重心	内積
リンク構造	リンク構造	重心	距離

3.1.　ベクトル基底の選択

(1) 文字列の場合

対象データが文字列の場合は，単語の出現を基底とするベクトルを作成する．ベクトル次元の増大を防ぐため，以下に示すTF・IDF法による手順により基底を決定する．

Step 1) 文書を索引語集合へ変換する．本稿では索引語をN=3のN-Gramにて抽出する．各要素が索引語集合である文書集合D'を得る．

Step 2) D'の全文書数をN，単語tのD'集合内における出現回数を$n(t)$，単語tを含む文書数を$d(t)$とおくと，索引語の重要度$w(t)$は次式のように求められる．

$$w(t)=\frac{n(t)}{N}\times\log\frac{N}{d(t)} \tag{1}$$

Step 3) 重要度$w(t)$の上位から任意数mの索引語tをベクトル基底として選択する．

Step4) 上記で得られたベクトル基底において索引語があれば1，なければ0の2値をとる要素を持つベクトルを作成する．

Step 5) ベクトルの長さが1となるよう正規化を行う．

(2) カテゴリカルデータの場合

対象データがカテゴリカルデータである場合は，以下の手順によりダミー変数を導入して数量データとなるようベクトル基底を決定する．

Step1) 文書集合内に出現するファイルタイプを抽出し，ファイルタイプ集合D'を得る．

Step2) ファイルタイプ集合D'内の要素を基底とし，基底が示すファイルタイプに文書が適合すれば1，適合しなければ0の2値をとる要素を持つベクトルを作成する．

Step3) ベクトルの長さが1となるよう正規化を行う．

(3) リンク構造の場合

対象データがリンク構造の場合は，次式に示す2つの要素から成るベクトルを構成する．

$$\mathbf{d}_j=[d_{1j}\quad d_{2j}]^t \tag{2}$$

d_{1j}：探索起点から文書jに到達するまでに辿るハイパーリンク数

d_{2j}：文書jが含むハイパーリンク数

3.2.　重心ベクトルによるフィルタの学習

あるトピックを示す次式で示される正例集合Dがあるときを考える．

$$\mathrm{D}=(\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_N) \tag{3}$$

ただし，$\mathbf{d}_j=[d_{1j}\quad d_{2j}\quad\ldots\quad d_{nj}]^t$

このとき，次式に示すDの重心ベクトル$\mathbf{q}$をトピックの特徴を表す代表ベクトルとする．

$$\mathbf{q}(\mathrm{D})=[q_1\quad q_2\quad\ldots\quad q_n]^t=\frac{1}{N}\sum_{j=1}^{N}\mathbf{d}_j \tag{4}$$

Information Filtering for Bidding Information Search System Based on Web Crawler

Naoyasu Omata†, Soichi Sekine†

KURIMOTO, LTD†

3.3. 類似度

比較の対象となるベクトル $\mathbf{d}, \mathbf{q}$ があるとき，類似度 $s(\mathbf{d}, \mathbf{q})$ を次のように定義する．

(1) 内積による類似度

次式に示すように内積を類似度とする．

$$s(\mathbf{d},\mathbf{q}) = \mathbf{d}\cdot\mathbf{q} = \sum_{i=1}^{n} d_i q_i \quad (5)$$

(2) 距離による類似度

次式に示すようにユークリッド距離 u を非類似度とする．

$$u(\mathbf{d},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(d_i - q_i)^2} \quad (6)$$

また，内積による類似度との比較のため，次式により値の範囲が $0<s(\mathbf{d},\mathbf{q})\leqq 1$ となるよう類似度を定義する．

$$s(\mathbf{d},\mathbf{q}) = e^{-u(\mathbf{d},\mathbf{q})} \quad (7)$$

(3) 正規化された類似度

複数の類似度を要素とする集合 S を得たときに，スコアの分布の平均値，標準偏差をそれぞれ μ，σ とおくと，分散が1となる正規化された類似度を得る．フィルタのスコアはこの正規化された類似度を用いる．

$$s'_j = \frac{s_j - \mu(\mathrm{S})}{\sigma(\mathrm{S})} \quad (8)$$

ただし，$\mathrm{S} = (s_1, s_2, \ldots, s_n)$

4. 入札情報フィルタの評価実験

4.1. 評価用データの作成

(1) Web ページの収集

大阪府内のある自治体を対象として，表 3に示す収集条件により Web ページの収集を行った．入札情報は一覧性のある目次ページよりリンクされているため，Web サイトを事前に調査し，深さ 2 以内で到達できる目次ページを探索起点とした．収集の結果として，8294 個のコンテンツを得た．

表3 収集条件

収集時期	2010年1月初旬
探索基点	目的情報に深さ2以内で到達できるWebページ
最大探索深さ	2

(2) フィルタ実験対象データ

発見したコンテンツのうち，テキスト化できるものを抽出し，キーワード「入札」でフィルタをした．結果として実験の対象となるドキュメント 815 個を得た．また，実験対象となったデータへ教師ラベルを付与する．該当自治体の入札公告の文書に該当すれば正例ラベルを，該当しなければ負例ラベルを付与した．ここで，正例の数は 100 個，負例の数は 715 個となった．

4.2. 実験手順

フィルタの訓練用サンプルの抽出

重心ベクトル $\mathbf{q}$ を求めるための文書集合を作成する．本稿では，正例ラベルが付与されたサンプルの半数からランダムに抽出した．ここで，訓練用サンプル 50 個を得る．

(1) フィルタの作成

重心ベクトルを求めフィルタを作成する．4つの評価観点に対応して4つのフィルタを得ることができる．ここで，4つのフィルタのスコアの相加平均をフィルタのスコアとする結合フィルタを作成する．したがって，実験の結果として得られるフィルタは5つとなる．

(2) スコアの計算

学習に使用していない残りの正例サンプル 50 個と，負例サンプル 715 個に対してフィルタをかけ，スコアを算出する．スコアの降順にソートし結果系列を得る．

4.3. 実験結果と考察

実験により得られた結合フィルタのスコア順位 10 毎の正例の出現率を示したものを図 3に示す．上位に正例が集中しており，良好な結果系列が得られていることがわかる．今回の実験では，理想的な結果系列を得られた場合は 50 位で累積正例出現率 100%を得ることになる．そこで，表 4に示すように 50 位まで見た際の累積正例出現率で比較を行う．結合フィルタと内容フィルタが最も高く 88%という高い水準である．URL フィルタ・ファイルタイプフィルタ・リンク構造フィルタは，低い水準内で同程度の性能となった．分析のため，フィルタから算出されるスコアの分布を図 4に示す．内容フィルタでは，スコア 1.5 付近を境界として，2つのクラスタが発生していることがわかる．そこで内容フィルタのスコア 1.5 以上の文書を確認すると全て正例であった．これは，正例集合内の入札公告文書は当該自治体で使用していると考えられる統一的な帳票であり，内容フィルタはその帳票であるかどうかの特徴をよく捉えているといえる．URL フィルタでは4つのクラスタが発生していることがわかる．最上位スコアのクラスタ内を見てみると入札情報が存在する Web サイト領域に共通の URL パスが得られていた．確かに類似は得られているといえるが，入札案件の説明として付属する文書も混在してしまうという結果となっていることが性能劣化の原因であった．リンク構造フィルタでは，リンク数が極端に少ないものであるか否かの2つのクラスタが発生している．これは，正例集合が全て PDF であったことが影響していると考えられる．上位のスコア 1.0 近傍のクラスタ内に含む文書は，他に表計算ファイルやワープロファイルを含む結果となった．ファイルタイプフィルタでは，PDF であるか否かを表す2つのクラスタが発生した．上位のクラスタには，負例集合内の PDF ファイルも全て含んだ結果となっている．

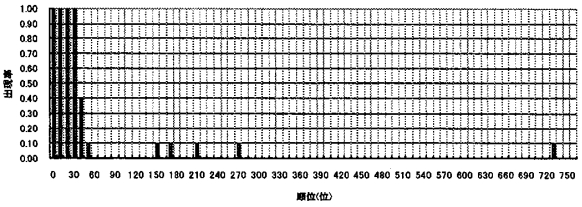
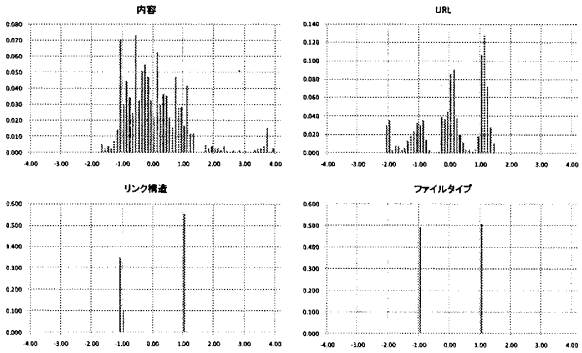


図3 結合フィルタによるランキング

表4 各フィルタ結果の50位までの正例累積出現率

フィルタ名	正例累積出現率
結合フィルタ	88%
内容フィルタ	88%
URLフィルタ	20%
ファイルタイプフィルタ	16%
リンク構造フィルタ	14%
※参考データ	
理想の場合	100%
フィルタしない場合	7.0%



※縦軸：頻度確率，横軸：類似度

図4 フィルタのスコア分布

5. 結論

自治体 Web サイトから得た文書集合に対して，複数の評価観点から成るフィルタリングを行った．結合フィルタ，内容フィルタでは良好な結果が得られた．入札情報フィルタを構成する上では，これらのフィルタを選択することが妥当である．その他の URL，ファイルタイプ，リンク構造では，フィルタを行わない場合に比べては良いが効果は低い．クラスタ発生の確認からヒントとなる情報は確かに得られるため，フィルタの結合方法の見直しを行うことによって，さらなる精度向上を図ることができると考える．

謝辞

本研究は（財）日本建設情報総合センターの研究助成を受けて実施したものです．

参考文献

[1] 小俣尚泰，関根聡一：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，クリモト技報，No.59，pp.46-55,2010.1