ISS-P-320

構造記述を活かした文書検索

小中　史人　　三浦　孝夫
法政大学理工学部創生科学科

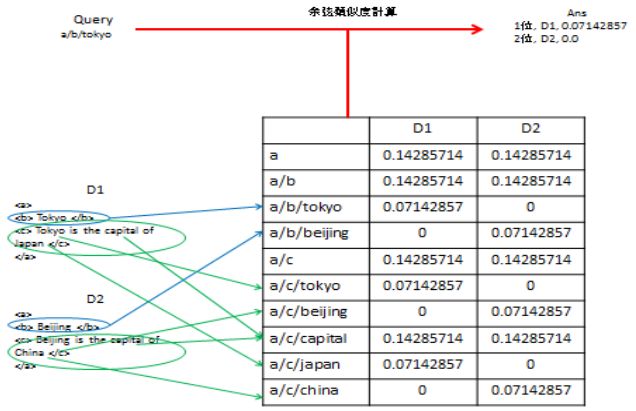
1.研究目的

インターネット上に点在する膨大な量の文書には、様ざまな種類が存在しそれぞれが特徴的な構造・構成を取ることが多い。これらを効率よく検索することで、大量の情報から所望する内容を的確に捉え、必要とする結果を得ることが可能となる。

本研究では､構造記述を持つXML形式の文書に対し、構造的特徴を活かした情報検索(IR)手法を提案する。このため、従来の Bag-of-Words 法によるベクトル化された補助情報の余弦類似検索に加え、タグ構造情報も対象にした 拡張 Bag-of-Words 法による余弦類似検索手法を提案する。

2.提案手法

一般的な手法では構造を考えず単語を全て同列に扱っているため区別することができない。本研究では、文書の構造ごとに単語を抽出し、不要語除去、ステミングを行う。その後、構造情報と処理した単語を纏めた上でベクトル化する。また、ベクトルの値としては TF×IDF の値を用いる。生成したベクトルは、元のコーパスに書き戻す。



	D1	D2
a	0.14285714	0.14285714
a/b	0.14285714	0.14285714
a/b/tokyo	0.07142857	0
a/b/beijing	0	0.07142857
a/c	0.14285714	0.14285714
a/c/tokyo	0.07142857	0
a/c/beijing	0	0.07142857
a/c/capital	0.14285714	0.14285714
a/c/japan	0.07142857	0
a/c/china	0	0.07142857

図.1　提案手法イメージ

図.1 は提案手法のデータの持ち方と検索のイメージ図である。まず、D1,D2 より構造 a, a/b, a/c を抽出する。その後 a/b に含まれる tokyo または beijing を抽出し、a/b/tokyo のように 1 つに纏める。a/c についても同様にする。ベクトルの値には TF×IDF 値を与え、コーパスに書き戻し、必要に応じて取り出す。

3.実験

本研究ではコーパスとして、販売されている Reuter Corpus の 1996 年 8 月 20 日の記事より名前昇順に 500 件を選択した。

比較手法として、単語情報のみを抽出したものに対して同様の処理を行って単語ベクトルを生成し､書き戻す。構造情報に関しては抽出したものを、頻度を回数としてベクトルを生成する。

今回は比較のために、クエリとの類似度を求める。類似度を計算する方法として、余弦類似度を用いる。比較手法では、クエリによって得られたそれぞれの結果を比較した上で再度ランクづけを行う。

評価基準はデータのタッチ回数の差とする。提案手法では、文書よりベクトルを取り出す際に 1、余弦類似度を参照する際に 1 とする。比較手法では、文書を参照し両ベクトルを取り出す際に 1、各余弦類似度を参照する際にそれぞれ 1、それらを照会する際に 1 とする。

検索条件	byline/ scott	dateline/ tokyo	title/ market	text/p/ market
提案手法	505	506	505	696
比較手法	1,153	6,127	3,505	100,699

図.2　タッチ回数

図.2 はいくつかの検索条件において、そのタッチ回数を比較したものである。検索条件は非一般的要求から一般的要求の順に並べている。

4.評価

図.2 より 2～145 倍の差があることがわかる。この理由としては、比較手法では、ほぼ全ての文書に出現している構造を要求した場合、要求単語を含む文書 1 つにつきそれら全てを参照しなければならない。それに対し、提案手法では、単語に直接構造情報を与えているため、別々に参照する必要がないからである。

5.結論

本研究では構造情報を持つ文書を、構造的特徴を捉えることで新たな検索方法を提案した。これにより検索要求を満たす文書を､従来の手法に比べ最大 0.7%のタッチ回数で抽出することができた。今後の課題としては、ベクトルを直接コーパスに書き加えているので、キーづけすることでデータサイズの膨張を抑えられるのではないかと考えられる。

参考文献

[1]　ALSAYED ALGERGAWAY, MARCO MESITI, RICHI NAYAK, GUNTER SAAKE：XML Data Clustering: An Overview, ACM Computing Surveys, Vol. 43, No. 4, Article 25, (2011)