Trend and Behavior Detection from Web Queries

Peiling Wang
Jennifer Bownas
Michael W. Berry

Overview

- WEBサーバログからマイニングできる
  クエリの特徴とその概要について

- 大学のWEBサイトに対する５０万クエリからの見地

Introduction

- WEB検索はエンドユーザに対しての検索を実現
  → WEB上の膨大な情報にアクセス可能

- とはいえ、ユーザの大半は上手ではない
  → 検索エンジンへの機能の追加は成功した
  → しかし、結局は使い慣れた人だけが利用可能

- 検索は、ユーザの検索方法をよりよく解釈する
  ことで有効性と使い勝手が更に向上するはず

Introduction(2)

伝統的なIRシステムでの問題指摘
(図書館のカタログ、オンラインデータベースなど)

1. クエリの文法的な複雑さ (難しさ)
2. ブール代数の記号の意味 (AND,ORなど)
3. 自然言語の言語的曖昧性

なので、色々と研究された。が…

色々と提案されているが、まだまだ余地がある

過去研究

・WEBエンジンでのクエリと検索セッションの
ログの結果からの解析

EXCITE（51473クエリ、1997年3月9日）
（1025910クエリ、1997年9月16日）

ALTAVISTA（993208159クエリ、
8月2日～9月13日、1998年）

Fireball検索エンジン（16，252，902クエリ、
ドイツにて、7月1日～7月31日、1998年）

過去研究(2)

どの研究も、データ収集と処理が違うが…
→　結構似てる

1. WEBクエリは非常に短い(平均2単語)
2. 単純な構造
3. 再利用されにくい
4. 文法的、意味的に間違いを含む
   →33%のゼロヒット

Term Associationの結果

全く違うパージング規則を使うExciteとAltaVista

語のペア
→ Zipf Frequency-ranking distribution**に従わない**

いくつかの頻度の高い単語ペアはトピックに関係なく
and-and, of-the, how-to, and-not, **など**

強いつながりの例：Cindy-crawford, visual-basic**のような、あるトピック単語の間では存在**

過去研究の問題点

・結局は短い期間のクエリでしかない
　→　WEB検索のごく一部のデータでしかない

もっと大きなデータ！！

大学のWEBサイトの検索機能による長期間
５４１，９２０のクエリのログ（1997年5月～2001年5月）

→長期的なトレンドと変化の分析を容易に！！

クエリデータとその分析

５４１，９２０クエリ（1997**年**5**月から**2001**年**5**月**）、
テネシー大のWeb**サイトの検索エンジンへのクエリ**

データマイニング（統計的、言語学的トレンド）
統計的分析　：　検索のグローバルな視野を提供

言語学的分析　：　主題、ボキャブラリと Term Association

Word Association：**単語の共起（隣接もしくは近傍）**
クエリの５４％は２～３語、６％がそれ以上

統計情報

全541，920クエリ

73，834　1997年
172，492　1998年
233，442　1999年　(増加！)
43，448　2000年
18，704　2001年　(減少！エンジン入替え、引退)

クエリの着目する特徴
ゼロヒット、空クエリ、先頭と末尾の文字の間での文字長、クエリ内での語数、そしてユニークもしくは一般的なクエリ、など

クエリ分析

- ゼロヒットクエリ：全体として３３％（３２％～４０％）
  -先行研究よりかなり大きい
    理由１：ストップワードがなし
    理由２：スペルミスが多い
    理由３：個別の名前を入力している
    理由４：ANDが基本
- 空クエリ（１．４％）
  EXCITE（5％）、ALTAVISTA（２０％）
- クエリの平均長　：　１３文字、最長１３１

クエリ分析（2）

- 非空クエリ：短い！
  40% 1語　　40% 2語
  14% 3語　　6% 4語、またはそれ以上
  最多語数：26語

- 平均2語、他の研究に比べてかなり短い
- 135，036クエリのうちの73%である98，236クエリが1度だけ出現
- 最も頻度の高いクエリ：“Career services”、9587回
- 2位：“Grade”、5727回
- 1000回以上のものは、すべてキャンパスライフに関連

Web検索のトレンド分析

図8．1
各月の傾向は同じ、ピークは1月で、6、8月は少ない

図8．2
ウィークデーはクエリが多い

クリスマスから新年までは殆ど検索が行われない
→大学の管理係が休みだから。。

自然対数による分析

図8．3（単語を対象）

・単語頻度のランク
（多くの単語が同じ頻度であることは不考慮）

・異なり頻度のランク
（頻度を基礎とする公平なランキングだが、語彙のサイズを不考慮）

→　6ぐらいまでは2本の線がオーバーラップ
7までにくっきりと分かれる

年を変えても傾向は同じ（図8．4）

自然対数による分析

単語ペアについては？（図8．5）
　→　単語のときと同様の傾向

・141353クエリのうち、65535クエリを利用
・49797ペアは2回以上出現

図8．3～8．5から、単語、もしくは単語ペアの、
頻度ーランクを1本の直線で表すのはムリ。

トレンドライン

図8．6：

- 2本の区分的な関数でトレンドラインを表現

- 2次関数（高頻度部分）と1次関数（低頻度部分）の組み合わせ（多項式）

- もう一本の線（異なり頻度）は2本の線では不十分
  → 低頻度には多くの単語（もしくはペア）のため

ボキャブラリの増大

クエリ数の増加＝＞異なり語数、単語ペアの増加

クエリ数の増加：はやい！
ボキャブラリ（異なり語）の増加：ゆっくり

ボキャブラリのオーバーラップ
5年すべてに出現：　2912
4年に出現　　：　2534
3年に出現　　：　3681
2年に出現　　：　6547
1年だけに出現　：　28168
（異なり単語数：44652）

単語ペアのオーバーラップ
5年すべてに出現：　2222
4年に出現　　：　3617
3年に出現　　：　7383
2年に出現　　：　17586
1年だけに出現　：　110545
（異なり単語ペア数：141353）

結論

- ボキャブラリは、Zipfの法則にあまり従わないが
  頻度ーランキングは、単語に基づくシンプルな多項式で
  規則性が見られる
- 共起（単語ペア）のマトリクスは、ボキャブラリ（単語）の
  一部だけが共起するためにスパース

課題！

- どうやって単語ペアを解釈するか？
- 単語のペアはANDが基本か？それともORか？
- 共起頻度が高いペアは個々の頻度が低くても
  高い重みを与えるべきか？
- 2つの単語はContent Descriptorになれるか？