Web コミュニティのリンク構造分析に基づく情報検索アルゴリズム*

久我 昌崇　中所 武司†
明治大学大学院 理工学研究科 基礎理工学専攻 情報科学系‡
{kuga, chusho}@cs.meiji.ac.jp

1 はじめに

インターネットにおける情報検索は、ディレクトリ型からロボット型、そして第３世代のリンク分析型へと変遷してきた。リンク分析型とは Google[1]、Clever Project[2] に代表される、Web のリンク構造に着目しサイトの評価を行なうものである。これらのシステムは従来のシステムより適合率が高いため注目されているが、それぞれキーワードの偏重、検索時間が長い等の問題点がある。

そこで、本稿ではその問題点を解決する新たな検索アルゴリズムを提案する。これは Web から検索トピックに関するコミュニティを抽出し、それに基づいて情報検索を行なう。

2 既存のリンク分析型検索システム

2.1 Google

Google は Stanford University の Brin,Page らによって開発された検索システムで、次のような特徴がある。

- リンクを他サイトへの推薦とみなし、backward-link（被リンク）の多いサイトに高ランクを与える
- ランクの高いサイトにリンクされているほどランクが高くなる

2.2 Clever Project

Clever Project は IBM Almaden-Research-Center で開発された検索システムで、次のような特徴がある。

- Google と異なり、リンクする側 (hub)、リンクされる側 (authority) の両面からサイトを評価する
- 有用な hub とは有用な authority を指すもので、有用な authority とは有用な hub に指されるものである。この再帰計算によりサイトを評価する

3 Web コミュニティ方式

3.1 Web コミュニティ

本システムではまず、検索トピックに関するコミュニティを抽出する。ここで Web コミュニティとは、共通のトピックを持つ集合体であり、同じ内容を持つサイトにリンクを張る傾向があることから Web では自然と形成されている。こういったコミュニティはそのトピックに関する専門家の集団とみなすことができる。そして、その専門家の意見 (リンク) を情報検索に利用するアルゴリズムである。

ここではコミュニティとしていわゆるリンク集を利用する。リンク集とはあるトピックに関する URL を多数載せているサイトであり、人手で作成されているので、リンク先サイトの質が高いことから検索に応用されることもある [3]。本システムではこのリンク集を利用して、適合率の高く、柔軟性のある検索を行なう。

3.2 アルゴリズム

具体的なアルゴリズムは次の通りである。(図１参照)

1. リンク集を求めるために入力キーワードを変形した query を作成する
2. query を既存の検索エンジンに送り、リンク集の集合を得る
3. そのリンク集の集合から forward-link の URL を抽出する
4. その URL 群を降順でソートする
5. 検索結果として返す

ここで図１の query = [入力 keyword * link * (URL=link) * (ANC=入力 keyword)] とはリンク集を得るために入力 keyword を変形したものである。keyword と文字列 link をテキストに含み、URL に文字列 link を含み、ANC(アンカー) に keyword を含む、という制限によりリンク集を得る。

3.3 システムの特徴

このシステムの特徴は以下の通りである。

- 要求トピックに関する専門家の集団 (コミュニティ) の意見 (リンク) を検索結果に反映することができるので検索結果の信頼性が高い
- キーワードの表記の揺れを吸収し、柔軟性のある検索を行なうことができる

*Information Retrieval based on Analizing Web community's Link structure

†Masataka KUGA and Takeshi CHUSHO
‡Computer Science Cource, Major in Sciences, Graduate School of Science and Technology, Meiji University.

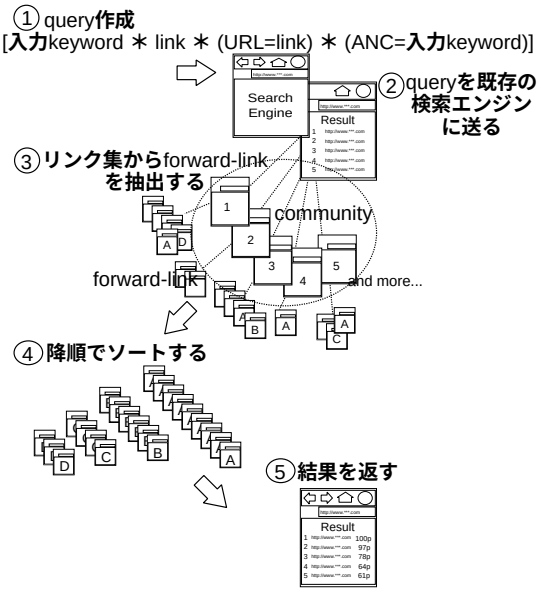

図 1: web コミュニティによる検索アルゴリズム図

4 比較・考察

4.1 Google との比較

Google のアルゴリズムは、テキストかそのサイトへのアンカーに入力キーワードを含むサイトしか結果として返さない。そのため適合率は高くなるが、検索に柔軟性がない。しかし、本システムでは同義語にも対応する柔軟性のある検索を行なうことができる。例えば、次のケースを考える。

---

user request = サッカーのセリエA情報が知りたい
correct answer = http://seriecalcio.8m.com
keyword = [serie * soccer]
Google's result … None in 48195 documents
keyword = [serie * calcio]
Google's result … 2th in 66196 documents

この例では seriecalcio.8m.com というサイトを１つの正解例としてある。なぜ、soccer と calcio という同じ意味を表すキーワードで検索したのに、一方では検索されずもう一方では６万を越える文書の中で上位２番目に結果として返されるのであろうか。これは Google が入力キーワードに完全一致する文書しか返さないからである。これではユーザは同義語についても調べない限り、良い検索結果を得ることができない。

こういった問題を解決する手法としてシソーラスを用いた同義語検索があるが、OR 検索を行なうと検索結果が大量になり、適合率が低下してしまう。

一方、本システムでは入力キーワードはコミュニティを求める際にのみ用い、そこから結果を求める際にはリンク情報を利用する。つまり、検索結果はコミュニティのリンク情報によるのであって、入力キーワードは直接的には関係しない。よって、キーワードに関連するがキーワードを含まないサイトを求めることができるので、結果的に同義語検索を行なえる。そして、リンク集という人手で作成された質の高いコミュニティを利用しているので、適合率も高い。例えば、下の表１はキーワードが soccer の場合の本方式の結果を示したものだが、この上位５件のうち３件は soccer という単語を含まない。つまり、１種類のキーワードで同義語検索を行なうことができ、適合率も低下しない。

このように web コミュニティ方式では、Google の検索キーワード偏重問題を解決することができる。

URL	被リンク数
http://nakata.net/	25
http://www.jfa.or.jp/	24
http://www.fifa.com/index.html	20
http://www.nidnet.com/link/fbj/	18
http://www.soccer.co.jp/ukiuki/	15

表 1: 本方式の結果 (keyword=soccer の場合)

4.2 Clever との比較

Clever のアルゴリズムは、まずルートセットと呼ぶ集合を既存の検索エンジンを用いて求めるが、その際に 200 回ほどの検索要求アクセスを行なわなければならない。これに対して、Web コミュニティ方式も既存の検索エンジンを用いるが、検索要求アクセスは１回で済む。このため、本システムでは Clever よりも高速な検索を可能としている。

5 おわりに

本稿では，Web からコミュニティを抽出しその知識を生かして検索を行なうシステムを提案した。そして本方式が Google に代表される既存システムの問題点を解決し、より柔軟性のある検索を行えることを示した。

参考文献

[1] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW7, Australia, Computer Networks 30(1-7) : pp.107-117, Apr. 1998.

[2] IBM Almaden Research Center, available at http://www.almaden.ibm.com/cs/k53/clever.html

[3] Sourav S. Bhowmick, Sanjay K. Madria, Wee-Keong Ng, Ee-Peng Lim, Web Bags-Are They Useful in A Web Warehouse?, FODO'98, Japan, Nov. 1998, available at http://www.cais.ntu.edu.sg:8000/tr/tr9813.ps