2008 年度卒業研究概要

Web 探索過程の可視化に有用な情報の蓄積に関する研究

大谷　紀子　研究室
0432203　宮本　憲一

1．研究の背景と目的

既存の検索エンジンはそれぞれ独自の方法でユーザが容易に情報を求めることができるように設計されているが、それぞれの検索エンジンは検索結果画面に Web ページの情報を羅列に表示したものがほとんどである。レポートや論文の作成など情報を体系的に得たい場合や情報の多様性をユーザが求める場合、検索エンジンから得た情報は結果の表示だけでは、ユーザはうまく情報を得ることができないため、探索過程を閲覧する方法が必要だと考えられる。探索過程を可視化する方法としては、トニー・プザンが提唱した図解表現方法であるマインドマップが有用だと考えられる。マインドマップとは、表現したい概念の中心となるキーワードやイメージを図の中央に置き、放射状にキーワードやイメージを繋げていくことで、発想を延ばしていく図解表現技法のひとつである。複雑な概念もコンパクトに表現でき、非常に早く理解できるとされる[1]。マインドマップの手法を参考にすれば、探索過程を可視化し放射状に伸びていく検索結果と探索過程を同時に表示する方法が可能になると考えられる。探索過程の可視化を実現するためには、既存の検索エンジンから得られる情報以外にも様々な情報が必要になる。

本研究では、Web 探索過程の可視化に有用な情報の蓄積を目的とする。Web 探索過程の可視化に必要だと考えられる情報を考察・提案し、提案した手法を評価するためのシステムを作成する。

2．情報の蓄積方法の提案およびシステム概要

Web で検索を行う際に使用する検索語句や、検索結果から得られる web ページの URL をより有効に活用することができれば、ユーザは体系的かつ多様性に富んだ情報を得ることができると考えられる。各検索語句から得た検索結果に含まれる URL の重複度の割合から、異なる検索語句同士の関連度を算出する。関連度を求める式を以下に示す。

関連度 = (重複した URL の総数)÷(検索結果から得た URL の総数)

また、検索結果から得られる URL のドメインの内訳を調べ、偏重値を算出する。偏重値とは、検索結果の URL をドメインによって体系立った情報、多様性のある情報に分け、その差により検索語句の偏重具合を示すものである。go,ac,or,gr など公的機関、研究機関、非営利団体のドメイン(以下、公的ドメイン)が含まれている機関が発信する web ページは、発信者によるバイアスが少ないため、体系立った情報として扱う。それ以外のドメインが含まれた web ページは、発信された情報に発信者によるバイアスが多いため、多様性のある情報として扱う。偏重値は以下の式によって算出する。

偏重値 = (公的ドメインが含まれる URL の総数)－(それ以外の URL の総数)

システムの動作の流れを以下に示す。

(1) 検索語句を入力する

(2) 入力した検索語句の DB が存在しない場合(3)へ、存在する場合(4)へ

(3) 入力された検索語句の DB を作成し、Google API から得た URL の内訳の情報を DB に保存する

(4) 入力された検索語句以外の検索語句との関連度を調べ DB に書き込む､ほかに検索語句がない場合は(1)に戻るかシステムを終了する

3．評価実験

システムの有用性を確認するため､検索から得られた上位 50 件の Web ページの URL を用いて評価実験を行なった。以下では、環境問題を a、地球温暖化を b と略記する。

表 1：評価実験 1 におけるドメインの内訳と偏重値

検索語句		a	b	a,b	a,b 対策	a,b 原因	a,b 現状
ドメイン	AC	3	0	0	0	2	4
	CO,COM	12.4	4.6	9.8	9.4	7.9	7.4
	GO	1	13	5	6	1	9
	OR	7	7	0	0	0	6
	NE	8	3	5	3	3	2
	GR	0	0	3	1	0	1
	perf,metro,city,vill,town	1	3	2	8	2	4
	pdf	0	0	1	1	1	15
偏重値		-13	9	-12	-1	-14	11

評価実験 1 において、各検索結果の URL からドメインの内訳を算出し、偏重値を求めた。表 1 は各検索結果の URL に含まれるドメインの内訳と偏重値を示したものである。検索語句[a,b,対策]のドメインの内訳は均衡が取れていたが、それ以外は偏りがあった。検索語句[a,b,原因]は公的ドメインの数が最も少なく、[a,b,現状]は PDF ファイルの数が最も多い結果になった。

評価実験 2 においては、検索語句[a,b,原因]に焦点を当て、新たに[a,b,原因,自然的要因]，[a,b,原因,人為的要因]，[a,b,原因,二酸化炭素]，[a,b,原因,気候変動]という検索語句を生成し、検索語句同士の重複度を算出した。[a,b,原因]に対する各検索語句の重複度はそれぞれ、0.28,0.20,0.52,0.30 であった。これにより、地球温暖化において自然的要因と人為的要因に大きな差はないことがわかった。また二酸化炭素のほうが気候変動よりも関心が高いことが分かった。

4．考察

評価実験1,2 より[a,b,対策]，[a,b,現状]がともに公的ドメインを多く含んでいるにもかかわらず 、[a,b,原因]は公的ドメインが少ない結果が得られた。通常、原因があって現状があり、対策を立てるということが筋道だと考えられるが、原因があいまいなまま現状の確認、対策を公的機関や民間が講じていると考えることができる。また二酸化炭素が地球温暖化の主原因のひとつである点、人為的要因、自然的要因の判別がついていない点が、重複度を求めることで分かった。このように、検索語句の関連性、URL に含まれるドメインの判別を行うだけで、情報探索の手がかりを見つけることが可能であると考えられる。今後の課題として、ドメインの判別をさらに詳細にすることで、より詳細な偏重を探ることが可能であると期待できる。また検索語句の関連性を算出する上で多様性のある検索語句の自動的な生成も今後必要だと考えられる。

参考文献

[1] トニー・ブザン著,田中孝顕訳,「人生に奇跡を起こすノート術―マインド・マップ放射思考」，きこ書房,2000.