의미 카테고리와 하이퍼링크를 이용한 검색엔진의 성능 향상

김형일[0] 김준태
동국대학교 컴퓨터공학과
{hikim, jkim}@dgu.edu

Performance Improvement of a Search Engine Using Semantic Category and Hyperlink

Hyung-Il Kim[0] Jun-Tae Kim
Dept. of Computer Engineering, Dongguk University

요 약

현재, 웹의 정보는 사용자들이 원하는 모든 정보를 담고 있다고 할 수 있으나, 방대한 웹에서 사용자가 원하는 정보를 정확히 추출하기란 어려운 문제이다. 이러한 정보 추출의 어려움은 방대한 정보량과 정보 추출 방식과 직결된다. 웹에서 정보를 정확히 추출하여도 일반적인 검색엔진들의 웹 페이지 순위 결정 방식을 따르게 되면, 사용자에게 중요한 페이지를 상위에 위치시키기란 쉬운 일이 아니다. 본 논문에서는 질의어의 모호성을 해결하기 위해 워드넷 기반 사용자 인터페이스를 설계하고, 웹 페이지의 가중치에 의미 카테고리 빈도 확률과 하이퍼링크 가중치를 이용한 웹 페이지의 가중치 결정 방식을 제안한다.

1. 서 론

인터넷의 급속한 성장은 시·공간의 제약을 극복할 수 있는 정보 가상공간을 만들게 되었으며, 이러한 환경은 정보의 집합 장소라고 대변될 수 있는 웹이라는 대규모 정보 저장 공간을 이루었다. 그러나 방대한 정보량은 선별의 어려움이라는 역기능을 창출하기도 하였으며, 이러한 정보의 역기능으로는 사용자에게 선택이라는 혼란을 발생시켰다. 이러한 정보검색에서의 정보 선택의 문제점으로 현재는 정보검색의 편리성이 강조되어 지고 있으며, 현재는 사용자 중심적 정보검색에 관한 연구가 활발히 진행되어 가고, 전통적 방식인 내용기반 방식의 문제점을 해결하기 위한 노력이 가속화 되고 있는 추세이다.

이러한 연구에 힘입어 차세대 검색엔진인 구글(Google), 다이렉트히트(DirectHit)는 사용자들에게 많은 호응을 얻고 있다. 차세대 검색엔진들은 전통적 방식인 내용기반 방식의 문제점들을 해결하여 사용자의 질의에 대해 적합한 웹 페이지를 추출하는 방식을 채택하고 있다. 그러나 이러한 시도에도 불구하고 검색 질의어의 다의성 문제로 인해 중요 페이지의 추출에는 여전히 어려움이 잔존하고 있다.

이러한 검색엔진에서의 문제점을 본 논문에서는 워드넷(WordNet)을 이용하여 질의어의 모호성을 해결하고, 웹 페이지 가중치 방식에 의미 카테고리 빈도 확률과 하이퍼링크(Hyperlink) 가중치를 이용하는 웹 페이지의 가중치를 결정 방식을 제안한다.

2. 관련연구

차세대 검색엔진 중 각광을 받고 있는 구글은 스탠포드대학에서 개발이 진행되어 현재는 상용화된 검색엔진으로 국내에서도 많은 각광을 받고 있다. 구글에서 이용한 웹 페이지 가중치 방식은 웹 페이지들의 하이퍼링크 정보를 이용하여 웹 페이지의 가중치를 결정하였다. 하이퍼링크 정보는 특정 웹 페이지 X가 웹 페이지 Y를 하이퍼링크로 가리키면 X라는 웹 페이지의 주제에 관해서 Y페이지는 중요도가 있다는 판단 하에 활용되었다. 이러한 하이퍼링크 정보의 활용은 Kleinberg의 HITS 알고리즘에 잘 소개되어 있다[5].

Kleinberg는 HITS 알고리즘은 in-link와 out-link를 활용하여 authority 페이지와 hub 페이지를 정의하여 웹 페이지의 가중치에 적용함으로 웹 페이지에 대표성(Representativeness)을 나타내었다[5]. 이러한 페이지의 구분 중, authority 페이지는 중요 정보를 많이 내포한 페이지라 할 수 있으며, hub 페이지는 중요 정보에 대한 링크를 많이 소유한 페이지라 할 수 있다. authority와 hub에 대한 가중치 공식은

$$a(p) := \sum_{q->p} h(q) \ , \ h(p) := \sum_{p->q} a(q)$$ 이다.

다른 방법으로는 명성 평가 기법이 있다. 명성 평가 기법은 주제에 관한 웹 페이지의 명성을 평가하는 방법으로, 웹 페이지가 얼마나 명성 있는가를 계산하는 기법이다[1]. 명성 평가 기법은 검색 질의에 대한 페이지의 순위를 측정하기 위해 penetration과 focus라는 두 가지 비율을 사용한다. 주제를 t라 하고 페이지를 p라고 할 때, 주제 t에 대한 페이지 p의 penetration은 페이지 p를 가리키며 주제 t인 페이지 수를 주제 t에 관한 전체 페이지의 수로 나눔으로 측정된다. 주제 t에 관한 페이지 p의 focus는 페이지 p를 가리키며 주제 t인 페이지의 수를 페이지 p를 가리키는 페이지의 수로 나누어 측정된다[1]. penetration은 주제 t인 임의의 페이지가 페이지 p를 가리킬 확률을 말하며, focus는 페이지 p를 가리키는 임의의 페이지가 주제 t일 확률을 말한다[1].

본 논문에서 질의어의 모호성 해결과 웹 페이지의 가중치 저장 방식에서 사용한 워드넷은 프린스톤대학에서 개발한 온라인 사전이다[2]. 워드넷은 어휘의 표현을 의미를 이용하여 어휘를 분류하였으며, 어휘의 의미관계를 동의, 반의, 상위, 하등으로 나타냄으로써 어휘의 의미 구조가 잘 나타난다[7][8]. 워드넷에서는 어휘의 의미 포함관계로 인해 계층적 구조를 갖는다. 이러한 워드넷의 명확한 어휘 분류 체계로 인해 문서 분류나 자연어처리 분야에서 많은 활용을 하고 있다[2][9][10].

정보통신부 대학기초연구지원사업에서 일부 지원되었음.

3. 실험용 검색엔진

3.1 질의어의 모호성 해결

웹 검색에 있어서 질의어가 단일 어휘로 이루어졌을 경우, 일반적인 검색엔진은 질의어의 어휘만을 고려하여 웹 페이지를 추출하므로 질의어에 모호성이 존재하는 경우에는 검색 결과의 정확성은 떨어지게 된다[34][38][39]. 예를 들어, 사용자가 " 자바" (의미: 섬)에 관해서 조사하고 싶을 경우, " 자바" 를 검색 질의어로 사용하게 되면 자바 섬에 대한 검색 결과를 얻기 힘들다. 이러한 결과의 발생은 현재 프로그래밍 언어로 각광을 받고 있는 " 자바" (의미: 프로그래밍언어)의 웹 페이지들이 높은 가중치를 가지고 있기 때문이다.

본 논문에서는 이러한 단일 질의어의 모호성을 해결하기 위해 설계한 사용자 인터페이스에서는 사용자가 질의어를 입력하면 의미 선택기가 활성화되어 워드넷에서 질의어의 모든 의미들을 추출하여 사용자에게 전송하고 사용자에게 의미 선택을 요구한다. 사용자는 인터페이스에 나타난 동의어, 상위어, 주석을 확인하여 질의에 합당한 의미를 선택한다. 선택된 질의어의 의미는 질의어 확장기로 전달되고, 질의어 확장기는 선택된 검색 질의어의 의미에 해당하는 주석 중에 명사만을 추출하여 원시 질의어와 결합하여 재조합 질의어를 완성함으로써 단일 질의어의 모호성 해결과 확장 질의어를 완성한다.

3.2웹 페이지 가중치 결정 방식

웹 페이지들은 단일한 카테고리에만 속하지 않는 경우가 빈번하게 발생되어 웹 페이지의 가중치를 단일한 결정 값으로 생성하는 것은 정보 활용 측면에서 위험한 문제를 야기할 수 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 웹 페이지의 가중치를 카테고리별로 저장하였다. 어휘의 의미에 대한 카테고리를 활용하여 웹 페이지에 나타낸 어휘들의 의미 카테고리 빈도를 분석하게 되면 해당 웹 페이지는 어휘의 의미들에 의해 몇 개의 카테고리에 수렴하게 된다.

본 논문에서는 웹 페이지에 나타난 어휘의 의미를 카테고리 빈도 확률로 이용하고 웹 페이지 가중치에 활용하고, 웹 페이지들의 연결 수단인 하이퍼링크를 이용한 웹 페이지의 가중치에 적용하는 웹 페이지 가중치 결정 방식을 제안한다. 실험용 검색엔진의 가중치 결정 방식은 웹 페이지의 가중치를 카테고리로 세분화하여 사용하고, 해당 웹 페이지를 가리키는 인링크의 총수를 가중치에 추가하여 가중치 결정 방식을 완성한다. 웹 페이지의 인링크 가중치와 카테고리 빈도 가중치를 웹 페이지의 가중치 결정에 절반씩 참여되도록 하였다.

웹 페이지에 나타난 어휘 의미들에 대한 카테고리 빈도를 SF_cat_i라 하고, 각 카테고리의 확률 빈도가 $Prob(SF\_cat_i)$ 이면 $Prob(SF_cat_i) = SF_cat_i / \sum_{i=1}^{n} SF_cat_i$가 된다. 이때, 최대 카테고리의 확률 빈도로 다른 카테고리 확률 빈도를 나누어 일반화하게 되면 카테고리 확률 빈도는

$log_2[((SF\_cat_i / \sum_{i=1}^{n} SF\_cat_i)/Max(SF\_cat_i/\sum_{i=1}^{n} SF\_cat_i))+1]$

이 되고, 결정된 카테고리의 확률 빈도는 웹 페이지 가중치에 활용된다. 임의의 페이지에 대한 인링크 가중치인 $P_i(Inlink\_w)$ 는 해당 웹 페이지를 가리키는 인링크의 총수 $Inlink_i_tot$에 대해 웹 페이지들을 가리키는 인링크 가중치들에서 가장 큰 값인 $Max(Inlink_i_tot)$로 나누어 일반화하면 $log_2[(Inlink_i_tot/Max(Inlink_i_tot))+1]$이 되고 임의의 웹 페이지의 가중치에 활용한다. 웹 페이지의 가중치 $Page_i_w$는

$log_2[((SF\_cat_i / \sum_{i=1}^{n} SF\_cat_i)/Max(SF\_cat_i/\sum_{i=1}^{n} SF\_cat_i))+1]$

+ $log_2[(Inlink_i_tot/Max(Inlink_i_tot))+1]$가 된다.

3.3워드넷 기반의 가중치 데이터베이스

대다수의 검색엔진에서는 가중치 데이터베이스를 특정 질의어에 대한 가중치 값과 해당 URL의 형식을 기초로 구성한다. 그러나 이러한 방식은 특정 질의어의 의미는 배제된 상태에서 가중치를 부여함으로 웹 페이지의 변별력을 감소시키는 원인으로 작용한다. 가중치의 변별력 감소에 대한 예제를 [표 1]에서 나타내었다.

검 색 어	URL_1의 가중치	URL_2의 가중치
Island (섬과 관련된 질의어)	50	10
Program Language(프로그램언어 관련 질의어)	10	10
Organization (단체에 관련한 질의어)	10	100
가중치의 총합	70	120

[표 1] 웹 페이지 가중치 방식에 대한 카테고리 분류 예제

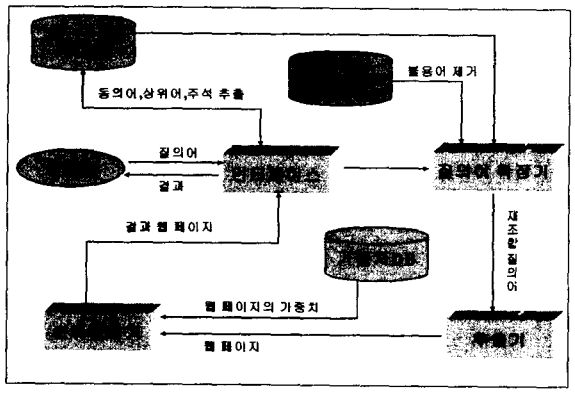
[표 1]에서 URL_1은 섬 관련 가중치로 50, 프로그램언어 관련 가중치로 10, 단체 관련 가중치로 10을 가지고 있다. 이러한 URL의 가중치가 저장되어 있을 때, 검색질의어로 JAVA (섬)가 사용되어진다면 일반 검색엔진에서는 URL_2의 가중치 총합이 가장 높기 때문에 결과 웹페이지로 URL_2가 나타나게 될 것이다. 그러나 위 표를 보면 해당 질의어의 의미에서는 URL_1이 URL_2보다 중요도가 높은 웹페이지이다. 이러한 가중치 방식의 문제점 해결과 질의어의 모호성 해결을 위해 본 논문에서는 워드넷을 활용한다.

아래 [표 2]은 본 논문에서 제시한 워드넷 기반의 웹 페이지의 가중치이고, 가중치 카테고리는 워드넷의 상위 카테고리 26개를 이용하였다.

URL_i (i = 1, 2, 3, ... , n) $CWi = Prob(SF\_cat_i) + P_i(Inlink\_w)$					
Category	C1	C2	C3	----	C26
Weighting	CW_1	CW_2	CW_3		CW_{26}

[표 2]웹 페이지의 가중치 형식

실험용 검색엔진의 시스템 구성도는 [그림 1]과 같다.



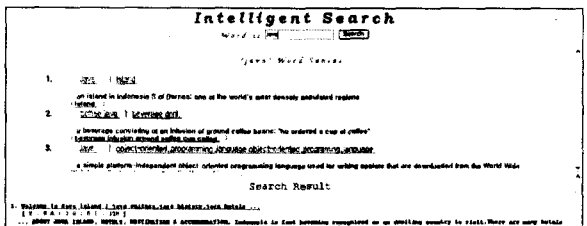
[그림 1]시스템 구성도

4. 실험 결과 및 분석

본 논문의 실험을 위해 질의어 당 200개의 페이지를 알타비스타에서 추출하여, 총 2,600개의 웹 페이지들을 실험 데이터로 결정하였다. 본 논문에서 사용한 하이퍼링크 가중치를 실험하게 위해 인링크의 총수가 필요하였으며, 웹 페이지의 인링크의 총수를 추출하기 위해 실험 전에 알타비스타를 활용하여 인링크의 총수를 추출하였다. 본 실험용 검색엔진의 성능 측정을 위하여 전통적 방식을 따르는 검색엔진인 알타비스타와 차세대 검색엔진으로 각광받고 있는 구글을 이용하여 비교 실험에 임하였다. 비교 실험에서는 결과 웹 페이지 중 상위 10개만을 추출하여 비교 실험에 사용하였으며, 웹 페이지의 관련도 판단을 위해 동국대학교 정보검색 전공자 6명을 선정하여 결과 웹 페이지에 대한 정확도 점수를 1부터 10까지 활용하여 정확도 점수를 부여하게 했다. 이렇게 나온 결과 점수들을 평균하여 7점 이상의 값을 부여 받는 웹 페이지들을 관련 있는 웹 페이지라 가정하였으며, [표 3]은 실험 결과이다.

사용된 질의어		일반 검색엔진				실험용 검색 엔진
어휘	의미	Altavista		Google		
		단일 질의어	복수 질의어	단일 질의어	복수 질의어	단일 질의어
Java	커피	0	5	1	7	8
Java	섬	0	4	0	5	7
Java	언어	6	8	7	6	8
Custom	통관	0	1	0	4	8
Custom	관습	2	2	1	4	5
Horse	마약	0	3	0	6	4
Horse	말	5	6	6	8	7
Coach	객차	0	4	1	6	3
Coach	코치	5	5	8	7	7
Plant	공장	2	6	4	8	6
Plant	식물	5	7	2	6	7
Sentence	문장	2	6	3	8	8
Sentence	판결	0	2	1	6	7
관련 문서의 개수		27	59	34	81	85
평 균 정 확 도		20.77%	45.38%	26.15%	62.31%	65.38%

[표 3] 실험 결과



[그림 2] 사용자 인터페이스

[표 3]의 실험 결과를 보면 일반 검색엔진에서 단일 질의어를 활용할 경우에는 질의어의 모호성으로 인해 관련도 높은 웹 페이지들이 사용자에게 전달되지 않았으나, 복수 질의어를 활용할 경우에는 질의어의 모호성이 다소 해결되어 단일 질의어를 활용한 경우보다 높은 정확도를 나타낸다. 실험용 검색엔진은 사용자의 의미 선택 행위로 단일 질의어에 대한 모호성이 해결되고 웹 페이지의 가중치 세분화로 인해 웹 페이지들에 대해한 변별력이 증가되어 일반 검색엔진보다 향상된 실험 결과를 나타낸다.

[그림 2]은 실험용 검색엔진의 사용자 인터페이스이며, A는 알타비스타의 가중치, G는 구글에서의 가중치, I는 실험용 검색엔진의 가중치를 나타낸다.

5. 결론 및 향후 연구 과제

본 논문에서는 질의어의 모호성 해결과 웹 페이지에 나타난 어휘의 의미를 이용하여 의미 카테고리 빈도 확률 가중치와 하이퍼링크 가중치를 활용한 웹 페이지의 가중치 결정 방법에 대해 제안한다. 본 논문에서 제안한 웹 페이지 가중치 결정 방식은 웹 페이지들에 대해 변별력을 증대시키게 되어 검색엔진에서 성능 향상을 도모할 수 있다. 향후 연구 과제로는 웹 페이지의 구조 분석을 이용한 웹 페이지 가중치 결정 방식에 대한 연구이다.

6. 참고문헌

[1]E. Agichtein, S. Lawrence, and L. Gravano. "Learning search engine specific query transformations for question answering", In Tenth International World Wide Web Conference, Hong Kong, 2001.5.

[2]C. Fellbaum, "WordNet : An Electronic Lexical Database", MIT Press,1998

[3]W. Frakes, and R. Yates, "Information Retrieval : Data Structures & Algorithm", Prentice-Hall,1992

[4]R. Hoch, "Using IR Techniques for text classification in document analysis", SIGIR'94, 1994

[5]J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", The Journal of the ACM, Volume 46, Issue 5, 1999.

[6]X. Li, S. Szpakowicz and S. Matwin, "A WordNet-based Algorithm for Word Sense Disambiguation", IJCAI-95, 1995

[7]G. A. Miller, "WordNet : An On-Line Lexical Database", International Journal of Lexicography, 1990

[8]S. Scott, and S. Matwin, "Text Classification Using WordNet Hypernyms", Coling-ACL '98Workshop, 1998

[9]M. Sanderson, "Word sense disambiguation and information retrieval", Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, 1994.

[10]E. Siegel, "Disambiguating Verbs with the WordNet Category of the Direct Object", Coling-ACL '98 workshop,1998

[11]E. Voorhees, " Using WordNet to disambiguate word senses for text retrieval" , Proceedings of the 16th ACM-SIGIR Conference, 1993.

[12]E. Voorhees, " Query expansion using lexical-semantic relations" , Proceedings of the 17th ACM-SIGIR Conference, 1994.

[13]E. M. Voohees, "Query Expansion Using Lexical-Semantic Relations", SIGIR'94, 1994

[14]WordNet, http://www.cogsci.princeton.edu/~wn/