DOI: 10.3745/KIPSTB.2008.15-B.5.487

사용자 적합성 피드백과 구루 평가 점수를 고려한 블로그 검색 방법

정 경 석[†] · 박 혁 로[††]

요 약

대부분의 웹 검색엔진은 문서의 적합도와 중요도를 함께 고려하는 순위화 방법을 사용한다. 문서의 적합도는 문서가 사용자의 검색의도를 만족시키는 정도이고, 중요도는 인기 있거나 양질의 내용을 포함하는 등 문서의 품질을 표시하는 정도라고 할 수 있다. 지금까지 웹 문서의 중요도를 평가하는 방법으로 가장 성공적인 것은 하이퍼링크 구조를 사용한 방법이다. 하지만 블로그의 경우, 해당 블로그를 작성한 블로거와 그 블로거가 소유하는 다른 문서들을 알 수 있기 때문에 문서의 중요도를 평가하는 다른 방법을 생각할 수 있다. 본 논문에서 제안하는 방법은 사용자의 북마크와 클릭를 이용하여 문서의 중요도를 계산하고, 그러한 문서 점수를 바탕으로 블로거의 구루점수를 계산한다. 마지막으로 문서를 순위화할 때 해당 문서를 작성한 구루의 구루 점수를 반영한다. 이렇게 되면 구루점수가 높은 구루 블로거의 문서들이 상위에 검색됨에 따라서 전반적으로 검색 품질이 개선될 수 있다. 블로그 문서를 대상으로 한 실험결과 제안하는 방법이 기존의 전통적인 웹 검색 성능과 비교하여 정답집합과의 연관성이 높음을 알 수 있었다.

키워드 : 사용자 적합성 피드백, 구루, 블로그, 검색, 랭킹, 북마크

Blog Search Method using User Relevance Feedback and Guru Estimation

KyungSeok Jeong[†] · Hyukro Park[††]

ABSTRACT

Most Web search engines use ranking methods that take both the relevancy and the importance of documents into consideration. The importance of a document denotes the degree of usefulness of the document to general users. One of the most successful methods for estimating the importance of a document has been Page-Rank algorithm which uses the hyperlink structure of the Web for the estimation. In this paper, we propose a new importance estimation algorithm for the blog environment. The proposed method, first, calculates the importance of each document using user's bookmark and click count. Then, the Guru point of a blogger is computed as the sum of all importance points of documents which he/she wrote. Finally, the guru points are reflected in document ranking again. Our experiments show that the proposed method has higher correlation coefficient than the traditional methods with respect to correct answers.

Keywords : user relevance feedback, guru, blog, search ranking, bookmarks

1. 서 론

초기의 웹에서, 검색엔진의 목표는 가능한 많은 웹 문서를 수집하는 것이었다. 하지만 수많은 웹 문서가 존재하는 지금은 사용자가 원하는 가치 있는 웹 문서 정확하게 검색하는 것이 더욱 중요하게 되었다.

최근, 문서와 질의의 유사도 만을 고려하는 키워드 기반 검색 방법[1-3]의 한계를 극복하기 위해, 추가된 정보를 근거로 하여 웹 문서의 중요도를 평가하여 검색에 활용하는 다양한 방법들이 제안되었다. 이러한 방법들은 크게 2가지로 구분될 수 있다. 첫 번째는 웹 문서의 하이퍼링크 구조를 사용하는 것이고[4-6], 또 하나는 검색결과에 대한 사용자의 반응을 이용하는 것이다[7,8]. 하이퍼링크 구조를 이용하는 방법은 웹 문서 검색에 있어서는 매우 유용할 수 있으나 블로그(blog) 검색에 있어서는 그 유용성이 제한적일 수 있다. 왜냐하면, 블로그 문서는 의미 있는 하이퍼링크를 많이 포함하기보다는 자신의 생각과 정보를 정리한 글들이 다수 존재하기 때문에, 이러한 문서들을 대상으로 페이지 랭크 값을 반영되기가 쉽지 않기 때문이다.

한편, 블로그 문서는 그 문서의 작성자인 블로거(blogger)에 대한 정보를 쉽게 알 수 있는 문서이다. 따라서 블로그 문서를 검색할 때, 해당 블로그 문서를 작성한 블로거의 중

† 정 회 원 : SK communication 검색플랫폼팀 과장
†† 종신회원 : 전남대학교 전자컴퓨터공학부 교수
논문접수 : 2008년 3월 28일
수 정 일 : 1차 2008년 5월 7일, 2차 2008년 6월 12일
심사완료 : 2008년 6월 16일

요도를 평가하여 이 정보를 문서 랭킹에 이용한다면 검색 품질을 보다 개선할 수도 있을 것이다. 즉 블로그 검색에 있어서 단순히 질의 단어를 많이 포함하는 블로그 문서만 찾아주는 것이 아니라, 양질의 많은 문서를 작성한 구루(Guru)들의 문서를 상위에 검색함으로써 검색 결과를 개선할 수 있다. 이러한 구루 블로거의 검색의 필요성은 사용자들이 번거롭지만 bloglines, rollyo[16,17] 등의 RSS(Really Simple Syndication) 리더기에 유명한 구루의 블로그를 일일이 등록해 주기적으로 찾는 현상을 보면 쉽게 알 수 있다.

이러한 관찰을 통하여, 본 논문에서는 블로그 검색을 위한 새로운 방법을 제안하고 실험한다. 본 논문에서 제안하는 방법에서는 먼저 사용자의 피드백(feedback)을 이용하여 블로거의 문서들을 평가하고, 그 문서들을 작성한 블로거의 구루 점수를 계산한다. 이 구루점수는 해당 블로거가 작성한 문서들의 랭킹에 다시 반영되어서 중요도가 높고 구루점수가 높은 블로거의 문서들을 상위로 랭킹 시킨다.

블로거의 구루점수를 평가하기 위한 데이터로 사용자의 피드백 정보가 사용된다. 최근에 웹 검색의 인기가 늘어나고, 사용 가능한 자원이 증가되면서 자동화된 잠재적 적합성 피드백의 이용이 정보검색 분야에서 연구되었다. 잠재적 적합성 피드백은 웹 문서의 적합도와 간접적으로 연관된 페이지 점유 시간, 클릭 같은 검색된 웹 문서에 대한 사용자의 행동들을 포함한다. 잠재적 적합성 피드백은 대량의 데이터를 보유할 수 있어 의미 있는 값을 찾는데 용이하지만, 명시적인 적합성 피드백에 비해, 많은 노이즈를 포함할 수 있고 그 많은 데이터를 처리하기 위해서는 많은 비용을 필요로 하게 된다. 그래서, 본 논문에서는 사용자의 북마크와 같은 명시적 적합성 피드백을 추가하여 사용자의 구루 점수를 계산한 후, 문서에 반영하는 방법을 제시한다.

2장에서는 최근에 연구된 관련 연구에 대해서 소개한다. 그리고, 3장에서는 우리가 제안하는 웹 검색 랭킹 방법에 대해서 소개하며 4장에서는 제안하는 방법의 실험 결과를 묘사하고자 한다.

2. 관련 연구

여러 연구에서 키워드와 다른 종류의 정보를 조합하여 키워드 매칭의 한계를 극복하여 웹 검색의 품질을 향상시키려고 노력하였다. 크게 두 가지로 나누면, 웹 하이퍼 링크 구조에 대한 정보[4-6,10]와 검색된 웹 문서상의 사용자의 행동에 대한 정보였다[7-9,11].

하이퍼링크는 웹에서 웹 문서에 접근하기 위해 사용된다. 웹의 하이퍼링크를 분석함으로써 어떠한 문서는 중요한 문서를 추측할 수 있게 되었다. 한 중요한 웹 문서는 많은 다른 문서들로부터 링크되기 때문이다. Kleinberg[5]는 하이퍼 구조를 분석함으로써 중요한 페이지를 구별하는 것을 제안 하였고, Google[4,5] 은 검색결과에 유사한 방법을 사용하였다.

중요한 페이지를 구별하는 정보로써 검색결과에 따른 사용자의 피드백 정보이다. 만약 많은 사람이 어떤 웹 페이지를 클릭했다면 그 페이지는 중요하거나 인기가 있다고 간주할 수 있다. Google Blog[18]의 경우, 하이퍼링크 구조뿐만이 아니라, 사용자의 잠재적 피드백 정보와 수 많은 HTML 기반 특성들을 사용하였다.

여러 연구 그룹에서는 사용자의 상호작용과 잠재적 자질들 사이의 관계를 평가하기도 하였으며[7,8], Teoma[11]의 한 부분인 DirectHit 검색 엔진은 이러한 정보를 검색 결과에 이용하였으며, Digg[12]는 사용자에 의해 과학 기사나 기술이 강조된 인기 커뮤니티 웹 사이트로써 추천 같은 명시적 적합성 피드백에 기반한 랭킹 시스템이다.

블로그 전문 검색으로 가장 유명한, Technaroti[19]를 보면, 페이지랭크와 유사한 방식으로 블로그들간의 링크 구조를 파악하여 중요한 문서를 구분하며, 북마크와 추천 같은 명시적 적합성 피드백을 이용한다.

3. 제안하는 방법

이 장에서는 문서의 적합도와 중요도를 평가하기 위한 자질들에 대한 설명과 랭킹 함수로써 다중 선형 회귀분석에 대하여 언급한다. 마지막으로 구루 점수를 평가하는 방법과 시스템 구조에 대해서도 기술한다.

3.1 자질

우리는 크게 검색 정확도를 향상시키기 위한 검색 만족도를 두 가지로 나누고, 그에 따른 자질들을 크게 세 가지 분류로 나누었다. 그리고, 범주에 따른 대표적인 자질들은 경험적으로 가장 일반적이고 강력한 자질을 선택하였다.

〈표 1〉 각 범주에서의 대표적 자질

검색 만족도	자질의 범주	대표적 자질
적합도	텍스트 기반	필드 별 가중치
중요도	사용자의 명시적 피드백	북마크 카운트
	사용자의 잠재적 피드백	클릭 카운트

3.1.1 필드 별 가중치

대부분의 현존 검색 엔진들은 코사인 유사도 뿐만이 아니라 필드 별 가중치를 사용하고 있다. 이 방법은 사용자의 질의에 따른 적합한 문서를 찾기 위한 것으로써, 키워드 기반 자질로써 가장 널리 사용되는 방법이며, 텍스트 기반 정보 중에서 가장 가중치를 높게 주는 자질로 알려져 있다[22]. 예를 들어, 질의어가 문서의 본문에서 출현했을 때 보다 제목에서 발견되었을 때 필드 별 가중치가 더 높게 부여된다.

이 것을 수식으로 표현하면, 질의어 q와 문서 d의 필드 별 가중치 FW는 문서의 필드 별 가중치의 합과 같다.

$$FW(d,q)=\sum_{i=0}^{n}(t_i \times w_i)$$

이 자질은 우리가 제한하는 방법과 비교하기 위한 베이스라인 시스템의 유일한 자질로써, t는 단어의 빈도수를 사용하지 않고, 1(질의어가 필드에서 존재)과 0(질의어가 필드에서 존재하지 않음)만을 사용하였고, w는 사람에 의한 경험적인 가중치 값이다.

3.1.2 북마크 카운트

북마크 카운트는 사용자의 명시적 적합성 피드백으로써, 문서의 요약정보 및 본문의 내용을 확인 후, 추천하거나 다시 찾을 만한 문서를 갈무리하는 목적으로 행해지는 사용자의 입력 형태이다.

위의 목적을 갖는 북마크는 검색 중요도와 상당히 밀접한 관계가 있어, Delicious[21]에서도 중요한 랭킹 자질로 사용되고 있다. 특히 블로그에서는 사용자들의 이용이 빈번함에 따라, 의미있는 데이터를 쉽게 얻을 수 있다는 장점을 갖고 있다.

우리는 이를 위해 따로 북마크를 할 수 있는 인터페이스를 구성 하였다. 또한, 북마크에 대한 사용자 참여도가 성능에 끼치는 영향을 정량적으로 평가하기 위해 북마크를 입력하는 사용자의 수를 차이를 두어 테스트 셋을 나누어 구축하였다.

3.1.3 클릭 카운트

클릭 카운트는 사용자의 잠재적 적합성 피드백으로 사용자들이 정보를 찾는 행위로 파생되는 대표적인 자질이다[8]. 하지만, 명시적 적합성 피드백에 비해, 많은 노이즈가 포함될 수 있는데, 문서의 중요도에 악영향을 미치는 대표적 클릭 패턴은 다음과 같다.

- 질의 키워드에 대한 사전 정보가 없어, 검색결과의 상위 문서를 근거 없이 일단 클릭하게 되는 경우
- 타이틀과 요약정보를 보고, 문서를 클릭했으나 적합하거나 중요하지 않는 내용일 경우
- 적합하고 중요한 문서가 더 있으나, 한 두 개의 문서를 통해서 충분히 정보를 얻어, 클릭하지 않는 경우

3.2 다중 선형 회귀분석

다중 선형 회귀분석(multiple linear regression analysis)은 여러 연구 그룹[13, 14, 15]에 의해 연구 되었다. 특히, [15]는 검색엔진 결과의 랭킹에서 각기 다른 자질들에 대하여 각각의 영향력을 평가하였다. 이 논문은 이진 분류 문제로 해석하여 로지스틱 회귀분석, support vector machine, binary classification tree 를 사용하여 학습 및 평가하였다.

본 실험에서는 위의 실험과 달리, 문서의 적합도를 이진값(적합/비적합)로 표현하여 학습하는 것보다 비교적 상세한 실수 값으로 사용하려 했고, 그에 따라 랭킹 함수를 다중 선형 회귀분석을 사용하였다. 이 분석 방법은 매우 이해하기 쉽고 직관적이며 성능도 다른 것에 비해 크게 떨어지지 않는다. 다중 선형 회귀분석이 사용된 문서의 랭킹 평가 함수 DS와 구루점수가 추가로 반영된 DS' 는 다음처럼 정의가 될 수 있다.

$$DS=\alpha+\beta\times FW+\delta\times CC+\varepsilon\times BC$$

$$DS'=\alpha+\beta\times FW+\delta\times CC+\varepsilon\times BC+\gamma\times GS$$

DS와 DS' 는 평가 함수의 반응 값이며, FW, CC, BC, GS 는 각각 필드 별 가중치 값, 클릭 카운트, 북마크 카운트, 구루 점수 이다.

3.3 구루 점수 평가

블로그의 구루 점수 GS는 그 블로거가 작성한 문서들의 값으로 계산된다. 작성한 문서의 수가 많거나 중요한 문서로 평가된 문서가 많을 경우, 그 사용자의 구루 점수는 커지게 된다.

$$GS=\frac{\sum_{i=1}^{n}DS_i}{\max GS},$$

maxGS는 블로거 중, 최고 구루 점수를 말하며, 결국 사용자의 점수 maxGS에 의한 상대적인 값인 0~1 사이의 값을 갖게 된다.

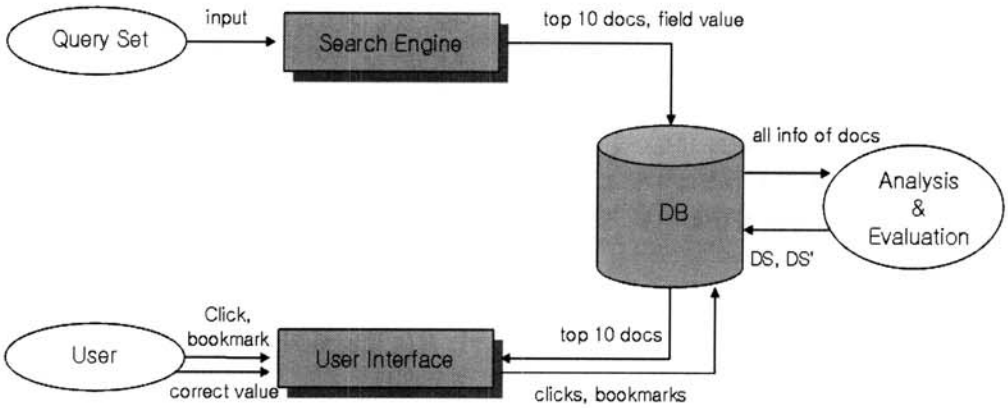
위의 구루 점수 GS 는, 그 구루의 모든 문서의 DS를 이용하여 계산되고, 이 GS 값은 다시 그 구루의 각각의 문서의 DS를 구하는 값을 보완하는 하나의 변수 값으로 사용되어, 최종적으로 구루 점수가 반영된 문서의 평가 함수 DS' 가 구해진다.

3.4 시스템 환경

제안하는 시스템은 우리의 랭킹 모델을 학습을 위해 데이터를 저장하는 구조로써 (그림 1)과 같다.

시스템은 크게 두 가지 모듈과 학습 및 평가를 위해, 모든 데이터를 저장하는 DB로 구성되어 있다. 첫 째, 검색엔진에서 질의어와 질의어에 따른 상위 10개의 검색된 결과와 필드별 가중치를 DB에 입력한다. 둘 째, 질의어 별 10개의 문서를 랜덤한 순위로 유저 인터페이스에서 브라우징 하고, 사용자들로부터 클릭과 북마크를 얻어 DB에 문서 별로 적재한다. 마지막으로, 지금까지 얻은 문서의 정보들을 다중 선형 회귀분석으로 문서의 랭킹 값을 구한 후, 그 문서를 기반으로 블로거의 점수를 계산하며, 이를 다시 그 사용자의 문서들에 반영하여, 각 문서 별 최종 랭크 값을 DB에 저장한다.

(그림 2)는 유저 인터페이스와 검색 샘플을 보여주고 있다. 유저 인터페이스는 클릭과 북마크를 얻는 것을 도와준다. 우리는 지정된 질의어 목록을 전문가에게 사용하도록 하였으며, 적합한 문서에 대해 자유롭게 클릭과 북마크를 하도록 하였다. 만약 사용자가 문서의 주제를 클릭하면 그 문서의 클릭 카운터는 1이 증가된다. 또한, "+플러스"를 누르면 북마크 카운트가 1이 증가된다.



(그림 1) 시스템 환경

(그림 2) 클릭과 북마크를 위한 유저 인터페이스

4. 실험 및 평가

4.1 데이터 셋

우리 실험에서는 북마크와 클릭 데이터를 쉽게 얻을 수 있는 블로그 문서를 확보하기 어려워 Trec과 같은 공신력있는 문서를 사용하지 않았고, 싸이월드[20]내의 블로그 서비스인 페이퍼로부터 랜덤하게 68만건의 문서를 수집하였다. 또한 질의어는 질의어 로그 중에 빈도가 높은 질의어를 1840개를 추출 하였다.

우리는 추출 문서와 질의어들에서, 의미 있는 질의어와 문서 쌍을 추출하였고, 최종적으로 질의어 1840개 중에서 1577개와 그 질의어들에 따른 결과 문서15810개를 선별하였다.

우리는 북마크의 입력율과 성능과의 연관 관계를 평가하기 위해 실험을 두 가지로 나누었다. 첫 번째 A 실험은 북마크 입력하는 사용자를 두 배로 늘려, 북마크 입력율을 높였고, 두 번째 B실험에서는 그에 비해 더 적은 사용자가 입력하도록 하였다.

〈표 2〉 질의어와 문서의 정보

Num	A	B	전체
모든 질의어 수	582	1318	1840
선택된 질의어 수	444	1133	1577
상위 10개의 문서	4817	10993	15810

〈표 3〉 학습 셋과 테스트 셋의 정보

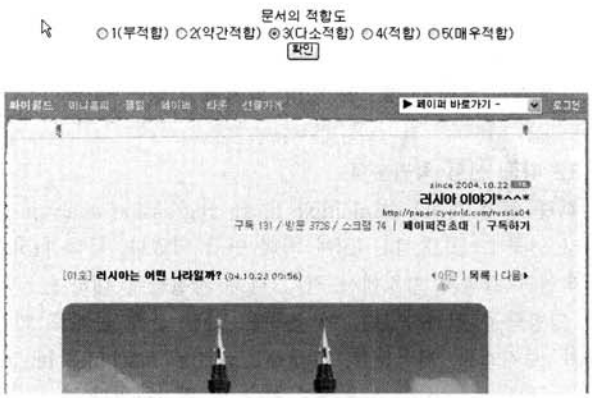
Set	A		B	
	질의어	문서	질의어	문서
테스트	360	84	924	219
학습	3907	910	8897	2096

〈표 4〉 BC와 CC의 입력율

	A		B	
	개수	비율	개수	비율
CC	657	68.1%	3497	78.6%
BC	308	31.9%	954	21.4%
Total	965	100%	4451	100%

<표 4>는 A실험과 B실험의 데이터 셋에서 *BC*와 *CC*의 입력율을 나타낸다. A실험은 B실험에 비해, BC가 CC와 비교하여, 10%정도 더 입력되었다.

정답집합과 정답집합의 점수는 북마크와 클릭을 사용하는 사용자들과 또 다른 사용자들에 의해서 5점 척도로 평가되었으며, 본문의 내용을 확인하여 평가할 수 있도록 아래와 같은 인터페이스를 구성하였다.



(그림 3) 정답집합에 대한 평가를 위한 인터페이스

4.2 평가 방법

우리의 평가 방법은 전통적인 평가방법인 precision, recall, f-measure를 사용하지 않는다. 랭킹 결과와 정답집합의 포함

관계로 평가하는 방식 보다, 랭킹 결과와 정답집합 문서의 순위 차이를 이용하여 유사성을 판단하는 지표를 얻고자 함이다.

본 평가 방법은 각 자질의 랭킹 값을 순위 값으로 변경하여 정답 셋 순위 분포와 비교하는 방법으로 비모수적 상관분석(nonparametric correlation analysis)을 사용한다. 비모수적 상관분석은 변수들이 순서 척도로 되어 있고 확률분포가 무엇인지 모르는 경우에 사용하는 방법으로, Spearman의 순위상관계수(Spearman's rank order correlation)와 Kendall의 타우(Kendall's tau)를 구하여 두 변수간의 상관관계를 조사할 수 있다. Spearman의 순위상관계수 r_s와 타우 t는 다음과 같다.

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

$$t = \frac{r_s}{\sqrt{\frac{1-r_s^2}{n-2}}}$$

d_i^2 : 두 변수 상의 순서값의 차이의 제곱
n : 표본의 크기

단, 유의 수준0.05를 기준으로, "자질 랭킹 값과 정답 셋 순위 사이에 연관성이 존재 하지 않는다"는 귀무가설을 기각할 수 없는 데이터일 경우, 이상치 데이터로 간주하고, 평가 셋에서 제외하였다. 최종 평균 상관계수 값 R(x)는 아래 수식과 같다.

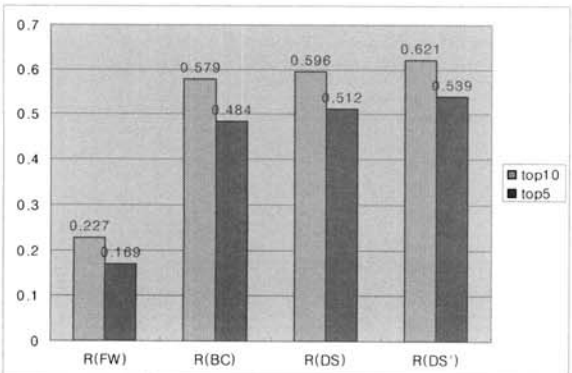
$$R(x) = \frac{\sum_{i=0}^{N} r_s}{N}$$

N : 평가 셋의 수

4.3 평가 결과

(그림 4)은 첫 번째 실험A의 결과를 보여준다. 필드별 가중치 *FW*과 북마크 카운트 *BC*, 다중 선형 회귀식인 *DS(=FW+BC), DS'(=FW+BC+GS)* 를 고려하였다.

*CC*는 *BC*와 상호 의존성이 매우 강하여, 다중 공선성(Multicolilinearity)이 발생되어 A실험 회귀분석 모델에서 제외되었다. 이 것은 사용자가 그 문서를 클릭 했을 경우, 대부분의 또 다른 사용자는 그 것을 북마크 했다는 것이다.



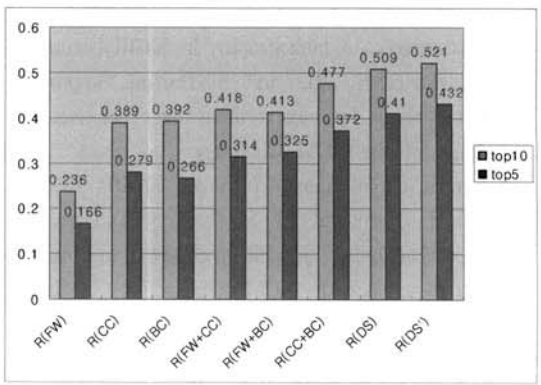
(그림 4) A실험 결과

우리가 제안한 구루 점수를 이용한 상관계수 값 *R(DS')*는 다른 것 보다 가장 높은 상관 값으로, top10일 때 0.621, top5일 때 0.539이었다. 이 수치는 정답집합과 상관관계가 비교적 강하다고 할 수 있으며, top5일 경우, *DS*의 상관계수 값 *R(DS)*에 비해 5.2%정도 향상 됨을 알 수 있어, 구루 평가 점수가 검색 성능 향상에 도움이 됨을 알 수 있다.

(그림 5)는 B실험의 결과를 보여준다. 필드별 가중치 *FW*과 클릭 카운트 *CC*, 북마크 카운트 *BC*, 다중 선형 회귀식인 *FW+CC, FW+BC, CC+BC, DS(=FW+CC+BC), DS'(=FW+CC+BC+GS)*를 고려하였다.

A실험과 달리, *CC*와 *BC*는 다중 공선성이 발생되지 않아 모두 B실험 회귀분석 모델에 모두 포함되었으며, 상관계수 값도 매우 유사하였다. 북마크의 입력율이 줄어들면서 클릭 역시 검색 품질을 향상시키는데 기여하는 것으로 분석된 것이다. 하지만 <표 4>를 보면 우리는 북마크의 입력율이 클릭의 21.4%로써, 더 적은 정보로도 비슷한 성능을 보였으므로 더 효과적인 자질임을 말 할 수 있다.

B실험에서도 기타 다른 방법들 보다 우리가 제안하는 *R(DS')*가 top10일 때 0.521, top5일 때 0.432로 가장 좋은 성능을 보였으며, R(DS)에 비해, top5인 경우, 상관계수 값이 5.3%정도 향상됨을 알 수 있다. B는 A의 실험에 비해, 전체적으로 상관계수 값이 0.1정도 낮은데, 이는 <표 4>처럼 BC 입력 비율이 10% 정도 낮기 때문이다.



(그림 5) B실험 결과

5. 결 론

본 논문에서 사용자 적합성 피드백과 사용자의 구루 평가를 이용하여 블로그 검색의 성능을 향상 시키는 방법을 제안하였다. 클릭, 북마크를 수집한 후, 랭킹함수를 도출하기 위해 다중 선형 회귀분석을 사용하였으며 이 랭킹 함수를 사용한 우리의 실험은 잠재적 자질보다 명시적 자질의 이용이 보다 효율적임을 밝혀냈고, 사용자의 구루 평가를 이용한 방법이 정답집합과 가장 관련도가 높음을 알 수 있었다.

*BC*의 입력율이 높아질수록 *CC*를 제외하여도 비슷한 성능을 보일 수 있음을 알 수 있었으며, *DS*에 구루 점수를 추

가한 *DS'*의 상관계수 값이 5.3% 더 높게 측정되어, 구루 평가의 사용으로 성능 향상이 있음을 확인 할 수 있었다. 또한 사용자의 참여나 블로그의 포스팅의 엄청난 증가 추이를 볼 때, 본 방법은 더욱 더 유용해 질 것이라 기대가 된다.

향후, 사용자 스팸성 피드백에 대한 연구와 스크랩이 많은 블로그의 특성 상, 스크랩된 문서와 실제 작성한 문서를 구분하여 보다 정교한 블로거의 구루 점수 찾는 방법을 향후에 과제로 남긴다.

참 고 문 헌

[1] R. Baeza-Yates and B. Ribiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[2] W. Frakes and R.Baeza-Yates, Information Retrieval: Data Structures & Algorithm (Prentice-Hall, 1992).

[3] G. Salton and M. McGill. Intorduction to modern information retrieval. McGraw-Hill, 1983.

[4] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, Proceedings of the 7th International World Wide Web Conference, 1998.

[5] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, The Journal of the ACM, Vol.46(5), 1999.

[6] B. Krishna and R. Monika, Improved algorithms for topic distillation in a hyperlinked environment, Proceedings of the 21st ACM SIGIR conference, 1998.

[7] D. Kelly and J. Teevan, Implicit feedback for inferring user preference: A bibliography. In SIGIR Forum, 2003.

[8] E. Agichtein, E. Brill and S. Dumais, Improving web search ranking by incorporating user behavior, In Proceedings of the ACM Conference on research and development, on information retrieval (SIGIR), 2006.

[9] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais and T. White, Evaluating implicit measures to improve the search experience, In ACM Transactions on Information Systems, 2005.

[10] Google, www.google.com

[11] Teoma, www.ask.com

[12] Digg, www.digg.com

[13] G. Pringle, L. Allison, and D. L. Dowe. What is a tall poppy among web pages?, Computer Net-works and ISDN Systems, 30:369-377, 1988.

[14] A. K. Sedigh and M. Roudaki. Identification of the dynamics of the google's ranking algorithm. In 13th IFAC Symposium On System Identification, 2003.

[15] A. Bifet and C. Castillo, An analysis of factors used in search engine ranking, First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[16] Bloglines, www.bloglines.com

[17] Rollyo, www.rollyo.com

[18] GoogleBlog, blogsearch.google.com

[19] Technorati, www.technorati.com

[20] Cyworld, www.cyworld.com

[21] Delicious, del.icio.us

[22] Seomoz, www.seomoz.org/article/search-ranking-factors

정 경 석

e-mail : humanist96@nate.com

2000년 조선대학교 전자계산학과(학사)

2002년 전남대학교 전산과(석사)

2004년 전남대학교 전산과 박사 수료

2004년~2006년 코리아 와이즈넛 연구소 연구원

2006년~현 재 SK communication 검색포탈본부 검색플랫폼팀 과장

관심분야 : 정보검색, 자연어처리, 음성합성

박 혁 로

e-mail : hyukro@chonnam.ac.kr

1987년 서울대학교 전산학과(학사)

1989년 한국과학기술원 전산학과(석사)

1997년 한국과학기술원 전산학과(박사)

1994년~1996년 연구개발정보센터 연구원

1997년~1998년 연구개발정보센터 선임연구원

1999년~현 재 전남대학교 전자컴퓨터공학부 교수

관심분야 : 정보검색, 자연어처리, 데이터베이스