블로그의 구조적 특성을 고려한 효율적인 블로그 검색 알고리즘

(The Effective Blog Search Algorithm based on the Structural Features in the Blogspace)

김 정 훈 † 윤 태 복 † 이 지 형 ††
(Jung-Hoon Kim) (Tae-Bok Yoon) (Jee-Hyong Lee)

요 약 오늘날, 대부분의 웹 페이지는 블로그영역에서 생성되고 기존의 웹 페이지 또한 블로그영역으로 전환되어가고 있다. 블로그 페이지는 트랙백연결, 블로거, 태그, 댓글과 같은 기존 웹 페이지에는 존재하지 않는 특징이 있다. 따라서 이러한 차이를 반영하지 않는 전통적인 웹 페이지 랭킹 알고리즘을 블로그 페이지에 단순히 적용하는 것은 효율적인 검색을 위해 적절하지 않다. 본 논문에서는 이러한 문제를 해결하기 위해 블로그 검색을 위한 "블로그-랭크" 알고리즘을 제안한다. 제안하는 알고리즘은 블로그의 구조적특징들을 활용하여 트랙백 연결성, 블로거의 명성, 사용자 반응성을 평가하고 이를 기반으로 블로그 페이지를 랭크 한다. 우리는 알고리즘의 검색효율성을 증명하기 위해 제안한 알고리즘을 적용한 블로그 검색 시스템을 구현하고 기존의 블로그 검색시스템과 검색효율성을 비교하였으며, 그 결과 블로그-랭크 알고리즘을 적용한 검색시스템이 기존의 검색시스템보다 더욱 뛰어난 검색효율성을 보임을 확인하였다.

키워드 : 랭크 알고리즘, 검색엔진, 연결성 분석, 블로그, 트랙백, 태그

Abstract Today, most web pages are being created in the blogspace or evolving into the blogspace. A blog entry (blog page) includes non-traditional features of Web pages, such as trackback links, bloggers' authority, tags, and comments. Thus, the traditional rank algorithms are not proper to evaluate blog entries because those algorithms do not consider the blog specific features. In this paper, a new algorithm called "Blog-Rank" is proposed. This algorithm ranks blog entries by calculating bloggers' reputation scores, trackback scores, and comment scores based on the features of the blog entries. This algorithm is also applied to searching for information related to the users' queries in the blogspace. The experiment shows that it finds the much more relevant information than the traditional ranking algorithms.

Key words : Rank algorithm, Search engine, Link-analysis, Blog, Trackback, Tag

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 연구 결과입니다. 연구비 지원에 감사드립니다(No. 2009-0075109).
· 이 논문은 제35회 추계학술대회에서 '효율적인 블로그 검색을 위한 블로그-랭크 알고리즘'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 성균관대학교 컴퓨터공학과
yeoshim@gmail.com
tbyoon@skku.edu
†† 종신회원 : 성균관대학교 정보통신공학부 교수
jhlee@ece.skku.ac.kr

논문접수 : 2008년 12월 19일
심사완료 : 2009년 6월 3일

1. 서 론

오늘날, 사용자 참여를 유도하는 웹 2.0 환경을 기반으로 대부분의 웹페이지는 블로그영역에서 생성되고 기존의 웹페이지 또한 블로그영역으로 전환되어 가고 있다. 따라서 블로그영역에 대한 효율적인 정보검색기술은 그 중요성이 날로 증가하고 있으며 그 중에서도 페이지 랭크 알고리즘은 이러한 정보검색기술의 가장 핵심적인 기술이다.

정보검색을 위해 사용자가 검색엔진에 질의하였을 때 검색엔진은 해당 질의어와 검색대상 페이지간의 연관성과 페이지의 사용자 만족도를 평가하고 그 평가를 기반으로 한 우선순위에 따라 결과리스트를 반환한다. 이러한 일련의 과정을 페이지랭크라 하며 이를 위한 알고리즘을 랭크 알고리즘이라 한다.

지금까지의 랭크 알고리즘 연구들을 살펴보면 크게 내용적인 측면과 구조적인 측면의 평가방법으로 나눌 수 있다. 내용적인 측면의 평가방법은 질의어와 관련된 단어들의 본문 출현 빈도수, 위치 및 특징과 같은 요소들을 기반으로 페이지의 내용을 직접 평가하는 방법으로 실용적인 결과를 얻기 위해서는 많은 계산 량이 요구된다. 반면, 구조적인 측면의 평가방법은 다른 페이지에 얼마만큼 많이 연결되어 있는지 혹은 좋은 페이지에 얼마나 많이 연결되어 있는지와 같은 연결성 평가를 기반으로 하므로 내용적 측면의 평가방법에 비해 훨씬 적은 계산 량을 필요로 한다. 따라서 이러한 구조적 평가방법이 매 순간 그 양이 급속도로 증가하는 오늘날의 웹 환경에서 새로운 내용을 실시간으로 갱신하는데 효율적이며 그로인해 검색의 정확성 및 효율성을 증가시킨다. 오늘날 가장 우수한 검색효율성을 보이는 PageRank[1]과 HITS[2]가 대표적인 구조적 측면에서의 평가방법이다.

이러한 기존의 웹 환경에 적용되었던 랭크알고리즘들을 블로그영역에 바로 적용할 수 있지만, 기존의 웹 데이터에 존재하지 않는 다음과 같은 블로그 페이지의 특징들을 고려한다면 더욱 효율적인 검색결과를 기대할 수 있을 것이다.

1. 블로그 페이지는 웹 페이지의 하이퍼링크와 같은 단순한 구조적 속성과 다른 트랙백연결, 태그, 댓글과 같은 구조적 속성들을 가진다[3]. 이러한 속성들은 블로그영역에서 활동 중인 블로거들에 의해 생성되고 수정되므로 그들의 생각, 관심, 반응들을 반영할 가능성이 높다. 또한 유사한 생각과 관심을 가진 블로거들과의 상호작용을 유도하여 더 많은 양질의 정보가 생성 및 관리된다[4]. 즉, 블로그영역의 정보는 이러한 구조적 속성들과 이들을 기반으로 한 상호작용을 통하여 점차 진화하는 것이다.
2. 하나의 블로그 사이트는 한명의 특정 블로거가 작성한 페이지들로 구성 및 관리되며 이로 인해 블로그는 흔히 "Personal publishing tool"로 불린다. 따라서 보통 블로그 페이지의 질과 주제는 그 사이트를 관리하는 블로거의 지식과 관심사항에 따라 결정된다. 이에 '블로거'라는 요소를 평가 및 관리하는 것은 더욱 효율적인 검색을 위한 중요한 하나의 요소가 될 것이다.

이와 같은 블로그영역의 특징들을 고려해 보면 블로그 페이지의 구조적 평가를 통하여 내용적 평가가 암묵적으로 수행된다는 것을 알 수 있다. 예를 들어, 블로거가 어떤 블로그 사이트에서 유용하고 관심 있는 페이지를 본다면, 자신의 사이트 이용자에게 그 페이지를 제공하고 싶어 할 것이고 자신의 사이트와 트랙백연결을 생성할 것이다. 혹은 트랙백연결을 통해 공유하지 않더라도 댓글을 통하여 자신의 의견을 페이지에 표현할 것이다. 즉, 트랙백연결과 댓글은 페이지의 내용이 누군가에게는 유용하고 관심이 있다는 것을 나타내는 하나의 요소라고 볼 수 있는 것이다. 결국, 이러한 속성들을 더 많이 혹은 더 좋은 것을 포함하는 페이지는 사용자들을 더욱 만족시킬 수 있는 정보가 될 가능성이 높다. 또한, '블로거'라는 가상적 요소를 평가할 수 있으므로 특정 주제에 대해 좋은 평가를 받은 블로거가 작성한 페이지는 좋은 내용을 가진 페이지일 가능성이 높다고 볼 수 있다. 즉, 페이지 내용을 직접 평가하지 않더라도 그 페이지의 트랙백 연결성, 사용자 반응성(댓글 기반) 그리고 블로거 평가를 기반으로 내용의 질을 예측할 수 있는 것이다. 이러한 현상은 페이지의 내용평가가 블로거와 사용자들의 블로그활동(블로깅)에 의해서 암묵적으로 수행되고 그 평가결과는 블로그영역의 구조적 속성에 자연스레 포함됨으로써 발생하는 것이다.

본 논문에서는 이러한 내용 평가를 반영하는 블로그 데이터의 구조적 속성을 기반으로 페이지를 평가 및 정렬하는 "블로그-랭크" 알고리즘을 제안한다. 제안한 알고리즘은 더 높은 트랙백 연결성과 사용자 반응성을 가진 페이지를 더 좋은 페이지로 평가하며, 또한 특정 주제에 대해 더 높은 명성을 가진 블로거가 작성한 페이지를 더 좋은 페이지로 평가한다. 이러한 제안을 실현하기 위해 본 논문에서는 다음과 같은 3가지 평가 요소를 기반으로 평가모델을 정의한다.

- 특정주제에 대한 블로거의 명성
- 페이지의 트랙백 연결성
- 페이지의 사용자 반응성

블로거의 명성은 트랙백연결, 태그, 그리고 댓글을 기반으로 다음의 두 가지 점수를 계산하여 평가한다.: (1) 특정주제에 대해 블로거가 작성한 페이지들의 평균점수, (2)블로그영역에서 특정주제에 대한 블로거의 활동점수. 페이지들의 평균점수는 태그, 트랙백연결, 그리고 댓글을 기반으로 계산하고 활동점수는 블로거가 작성한 좋은 페이지들의 작성횟수에 한 페이지가 기여하는 트랙백연결 가치를 가중치로 적용하여 계산한다. 트랙백 연결성은 트랙백연결의 수와 그것을 생성한 블로거의 명성평가를 기반으로 계산한다. 사용자 반응성은 댓글의 수에 하나의 댓글이 기여하는 트랙백연결가치를 가중치로 적용하여 계산한다.

본 논문은 또한 제안한 알고리즘을 적용한 검색시스템을 구현하여 기존의 블로그 검색시스템과 검색성능을 비교 하였다. 실험을 위해 티스토리 도메인[5]에서 200개의 블로그 사이트, 62906개의 페이지를 수집하였다.

제안한 알고리즘을 적용한 우리의 검색시스템이 기존의 티스토리 검색시스템보다 사용자 질의에 더욱 관련 있는 정보를 검색하였다. 이것은 블로그의 구조적 특징이 블로그검색의 성능과 유용성을 개선할 수 있다는 것을 보여준다.

본 논문의 구성은 2장에서 본 연구의 문제를 정의하고 목적을 더욱 명확히 한다. 3장에서 제안하는 블로그-랭크 알고리즘을 소개하고 4장과 5장에서 각각 실험설정과 실험결과에 대해서 기술한다. 마지막으로 6장에서 본 연구의 결론을 짓고 마무리한다.

2. 관련연구

지금까지 블로그 페이지 랭크를 위한 몇몇 연구가 있었다. Adar et al.[6]는 iRank라는 랭크 방법을 제안했으며 이 방법은 정보의 출처를 포함하는 페이지를 높게 평가하는 반면 블로그-랭크 알고리즘은 사용자들의 관심과 상호작용을 기반으로 사용자들에게 더욱 유명한 페이지를 높게 평가하는 방법이다. 이것은 최신 주제를 주로 다루는 블로그영역의 특성 상 사용자들에게 더욱 관심이 높은 유명한 페이지를 더 중요시하는 제안 알고리즘이 iRank보다 사용자에게 더욱 만족스런 정보를 제공한다. 본 연구와 유사한 Fujimura et al.[7]의 연구는 agent와 object간의 연결성을 EigenRumor vector로 정의하고 이 vector계산을 기반으로 페이지를 평가하는 방법이다. 이 방법은 정보의 제공성과 만족도 평가에 의존하는 방법으로 트랙백 연결성을 고려하지 않는다. 반면, 블로그-랭크 알고리즘의 사용자의 참여에 의한 트랙백 연결성 평가에 기반을 두고 또한 트랙백연결 연결은 페이지의 내용평가를 암묵적으로 반영하여 더 좋은 정보검색을 가능하게 한다. 블로그-랭크 알고리즘이 구조적측면의 평가방법이라는 점에서 PageRank, HITS와 유사하다. 하지만 블로그-랭크 알고리즘은 블로거의 명성점수를 평가 및 관리하므로 연결성 평가를 위해 기존의 알고리즘에서 요구되는 연결구성시간이 필요하지 않다. 즉, 블로거의 명성평가를 기반으로 연결성을 거의 가지지 않는 페이지 작성초기에 페이지의 효율적인 평가가 가능하다. Mishne[8,9]은 2006 TREC Blog Track에서 opinion retrieval의 효율을 개선하기 위한 방법을 제안하였다. 제안한 방법은 블로그의 속성들 중 작성시간, 댓글의 양 그리고 질의어에 대한 가중치를 활용하여 최근에 이슈가 되었던 페이지에 대한 검색효율성을 향상시켰으며 페이지에 포함된 opinion의 정도를 파악하였다. 이 방법은 블로그의 단편적인 정보들(시간, 댓글 양, 질의어 로그) 만을 활용한 방법인데 반해 블로그-랭크 알고리즘은 사용자의 평가와 의미적 연결성 평가를 고려하므로 더 나은 검색결과를 기대할 수 있다. Liu et al.[10]는 블로그 페이지의 랭킹을 위해 documentRelevance, documentFreshness, typeCoherence의 3가지 요소를 cosine 유사도로 계산하였다. documentRelevance는 질의어와 페이지 내의 단어의 출현빈도를 기반으로 계산하였으며, documentF- reshness는 페이지의 작성일자와 질의어의 발생시각의 차이를 기반으로 계산하였다. 마지막으로 typeCoherence는 사전에 질의어 분류를 사용자의 의도별로 분류해 두고 검색사용자의 질의어를 이 분류와 비교하여 사용자의 의도를 파악하는 방법이다. 이 방법은 단어 출현빈도 파악을 위해 페이지의 내용을 직접 파악해야 하며 사용자의 의도파악을 위한 질의어 분류에 사전/추가 작업이 필요하다. 이에 반해 블로그-랭크 알고리즘은 사용자의 자연스런 블로그활동(블로깅)에 따른 구조적 속성에 기반을 둔 평가방법이므로 직접적인 페이지의 내용파악이나 추가적인 작업이 필요 없다. 이로 인해 급속도로 증가하는 웹 상의 정보를 즉각 반영하여 우수한 검색결과를 기대할 수 있다. Java et al.[11]는 2006 Blog Track Open Task: Spam Blog Classification[12]에서 BlogVox라는 시스템을 제안하였다. 이 시스템은 유용한 내용이 아닌 광고만을 포함하거나 연결 수 증가를 위한 스팸 블로그 페이지를 검색 및 제거하는 시스템이다. 시스템은 SVM을 기반으로 질의어와 페이지 주제 간의 연관성을 파악하고 페이지 내에 주제와 관련된 opinion이 존재하는지 분석한다.

3. 문제정의

블로그 페이지는 기존의 웹 페이지와 다른 특성들을 가지며 그림 1은 이러한 블로그영역의 구조적 특징을 보여준다. 블로그 영역은 많은 수의 블로그 사이트들을 가진다. 하나의 블로그 사이트는 한명의 블로거와 그가 작성한 페이지들로 구성된다. 하나의 페이지는 내용, 태그, 트랙백연결, 댓글, 페이지를 작성한 블로거의 ID, 작성시간 등등으로 구성된다. 또한 블로그영역에는 블로그 사이트 간 두 종류의 상호작용이 존재한다.: 트랙백연결, 댓글. 트랙백연결은 블로그 사이트들의 페이지간의 상호작용이다. 댓글은 블로거와 다른 블로거 사이트 페이지의 상호작용이다.

블로그영역에서 활동하는 사람들은 크게 블로거와 일반사용자로 나뉠 수 있다. 블로거는 자신의 블로그사이트의 페이지들을 작성 및 관리하고 다른 사이트를 탐색하며 블로그연결과 댓글을 생성하는 사용자이고 일반사용자는 블로그영역을 탐색하며 댓글만을 생성하는 사용자이다. 블로거들은 자신만의 블로그사이트를 운영하며 이로 인해 그 사이트의 내용과 질은 그것을 운영하는 블로거의 관심과 능력에 따라 좌우된다. 또한, 그들의

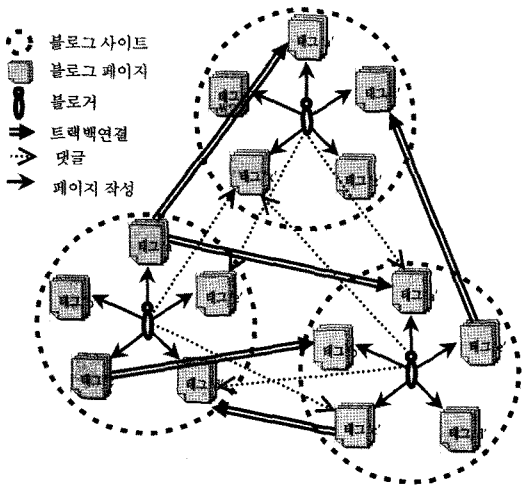


그림 1 블로그 영역

활동으로 인해 블로그영역의 페이지들은 그들 간의 연결성과 정보량을 점점 증가시켜 점차 진화하고 발전한다[3]. 따라서 본 논문에서는 페이지의 사용자만족성에 더욱 관련이 있는 블로거들을 주된 평가대상으로 하고 일반사용자는 댓글을 기반으로 하는 사용자 반응성 평가에 활용된다. 또한, 블로그영역의 페이지는 트랙백연결, 댓글과 같은 구조적 속성을 가진다. 블로거는 블로그영역에서 유용하다고 판단한 다른 블로거의 페이지를 자신의 사용자들과 공유하고 싶을 때 주로 트랙백연결을 달고 자신의 의견이나 생각을 그 페이지에 표현하고 싶을 때 댓글을 단다. 즉, 트랙백연결과 댓글이 많은 페이지는 사용자들에게 유용하거나 관심이 높은 페이지라 볼 수 있는 것이다. 따라서 본 논문에서는 블로그영역의 구조 중 블로거, 트랙백연결, 댓글을 페이지의 사용자 만족도에 영향을 주는 요소로 고려하며 아래의 가정에 따라 사용자에게 유용한 페이지를 판단한다.

- 더 좋은 블로거가 작성한 페이지가 더 좋은 페이지
- 더 좋은 블로거에 의해 생성된 트랙백연결이 더 많을수록 더 좋은 페이지
- 더 많은 댓글을 포함하는 페이지가 더 좋은 페이지

이러한 요소들을 고려하여 페이지의 점수를 다음과 같이 평가한다.

$$\text{페이지의 점수} = \text{블로거의 점수} + \text{트랙백의 점수} + \text{댓글의 점수} \quad (1)$$

또한, 일반적으로 사용자들은 웹에서 검색 시 보통 키워드를 사용하며 검색엔진은 그 키워드와 관련된 정보의 연관성을 평가하여 사용자에게 제공한다. 이러한 연관성평가를 위해 기존의 연구와 알고리즘들은 페이지 내용에 포함된 단어의 빈도수, 위치, 특징 등을 고려하였다. 하지만 이러한 방법들은 비싼 컴퓨팅비용이 필요하여 데이터가 매우 빠른 속도로 증가하는 실제 웹 환경에 적용하기에는 문제가 많다. 반면, 블로그영역은 '태그 메커니즘'이라는 구조를 가지고 있다. 이 구조를 활용하면 사용자의 질의어와 페이지 내용의 연관성을 보다 쉽게 파악할 수 있다. 보통 페이지를 작성한 블로거는 자신의 페이지 내용을 잘 나타내기 위해 태그를 신중히 선택하는 경향이 있다. 태그가 잘 정의된 페이지는 관리가 용이하며 검색의 효율성을 증가시키므로 이러한 페이지들이 더 많이 존재하는 블로그 사이트는 더 많은 사용자들이 관심을 가지는 사이트가 되기 때문이다. 즉, 페이지 내용과 태그와의 관련성 평가를 작성한 사람(블로거)에게 위임하는 것이다. 따라서 태그들은 그것들이 포함된 블로그 페이지의 내용과 관련이 있는 키워드들로 구성될 확률이 높다고 볼 수 있는 것이며 본 논문에서는 사용자의 질의어와 페이지 내용의 연관성 평가를 위해 태그를 활용한다. 즉, 키워드 k가 페이지에 태그로 존재한다면, 그 페이지의 내용은 질의어 k와 관련이 있을 가능성이 높은 것이다. 하지만 특정주제에 대하여 블로거 A와 블로거 B의 전문성은 차이가 있다. 예를 들어 전산학분야에서 "온톨로지"라는 키워드에 대해 지능시스템 전공교수와 인문학전공 대학생의 전문성은 차이가 있으며, 일반적인 관점에서 전자가 작성하는 페이지의 내용이 후자가 작성하는 페이지의 내용보다 더 양질의 데이터가 될 가능성이 높다. 따라서 키워드와 내용의 연관성평가를 위해 태그 메커니즘을 활용하기 위해서는 특정주제에 대한 블로거의 평가가 필요하다. 이러한 점을 고려하여 평가 식 (1)을 다음과 같이 수정한다.

$$\text{태그 t를 가진 페이지의 점수} = \text{키워드 t에 대한 블로거의 명성점수} + \text{트랙백의 점수} + \text{댓글의 점수} \quad (2)$$

이렇게 수정된 식 (2)를 살펴보면, 질의어 t가 질의되었을 때, 태그 t를 가진 페이지를 평가하기 위해 키워드 t에 대한 블로거의 명성점수와 해당 페이지의 트랙백과 댓글의 점수를 평가한다. 각 요소들의 평가를 위한 알고리즘은 다음 장에서 설명한다.

4. 블로그랭크 알고리즘

3장의 평가 식 (2)에 따라, 블로그-랭크 알고리즘을 다음과 같이 정의한다. 태그 t의 관점에서 페이지 e의

평가, $ES(e, t)$는

$$ES(e,t) = \begin{cases} 0 & \text{, if } e \text{ does not have tag } t \\ BS(b,t) + TBS(e,t) + CS(e,t) & \text{, otherwise} \end{cases} \quad (3)$$

where,

b: 태그 t를 포함하는 페이지 e를 작성한 블로거,

$BS(b, t)$: 태그 t에 대한 블로거 b의 명성점수,

$TBS(e, t)$: 태그 t의 관점에서, 페이지 e에 포함된 트랙백연결의 점수,

$CS(e, t)$: 태그 t의 관점에서, 페이지 e의 댓글 점수.

3장에서 정의한 페이지의 평가요소 태그, 블로거(작성자), 트랙백연결, 댓글을 기반으로 식 (3)은 페이지를 평가한다. 첫 번째 요소 $BS(b, t)$는 사용자(블로거)의 평가를 기반으로 페이지의 내용과 질의어 t의 연관성을 나타내고, $TBS(e, t)$는 페이지 e의 트랙백 연결성을 나타낸다. 마지막으로 $CS(e, t)$는 페이지 e의 사용자반응성을 나타낸다. 각 요소들은 다음 장에서 상세히 살펴본다.

4.1 블로거 점수(Blogger Score)

이 점수는 특정 주제(태그)에 대한 블로거의 명성을 나타낸다. 블로그영역에서 하나의 주제에 대한 블로거의 명성은 그의 과거 행동과 산출물에 기반을 둔다. 왜냐하면, 3장에서 언급하였듯이, 블로그영역에서 블로거의 기록 가능한 요소들 중 페이지의 작성(산출물), 트랙백연결(행동), 댓글작성(행동)은 페이지의 내용과 질을 반영하기 때문이다. 다시 말해, 만약 블로거가 한 주제(태그)에 대해 블로그영역에서 활발히 활동한다면, 그리고 그 활동의 산출물이 좋다면, 그의 명성은 높을 것이다. 이것을 위해 두 가지 관점의 점수를 고려한다.: 특정 태그에 대해 블로거가 작성한 모든 페이지의 평균점수(BES), 특정 태그에 대해 블로그영역에서의 블로거 활동점수. 한 점수가 블로거 점수에 크게 영향을 미치는 것을 방지하기 위해 두 점수를 정규화 한다. 이러한 사항들을 기반으로 블로거 점수는 다음과 같이 정의한다.

$$BS(b,t) = sigmoid(BES(b,t)) + sigmoid(BAS(b,t))$$
$$where,\ sigmoid(a) = 1 / (1 + e^{-a}) \quad (4)$$

1. 블로거의 과거 페이지 점수(Blogger Entries Score, BES)

이 점수는 특정 태그에 대해 블로거가 작성한 페이지의 질을 나타낸다. 본 논문에서는 블로거의 페이지 점수를 과거 블로거가 작성했던 페이지들(산출물)의 평균점수로 정의한다. 왜냐하면, 블로그영역에서 작성했던 페이지는 블로거의 생각과 관심을 반영하기 때문이다.

$$BES(b,t) = \frac{\sum_{e \in E_t^b} PES(e,t)}{|E_t^b|} \quad (5)$$

where, E^b_t : 블로거 b가 태그 t에 대해 작성한 모든 페이지의 집합

본 논문에서 블로거가 과거에 작성했던 페이지는 블로거의 명성이 없는 상태에서 작성한 페이지라 가정한다. 따라서 페이지의 점수계산식(3)에서 블로거 점수(BS)를 기본 값 1로 설정하여 과거 작성한 페이지의 점수(PES)는 다음과 같이 정의될 수 있다.

$$PES(e,t) = 1 + TBS(e,t) + CS(e,t) \quad (6)$$

2. 블로거의 활동 점수(Blogger Activity Score, BAS)

이 점수는 특정주제(태그)에 대해 블로그영역에서의 블로거 활동성을 나타낸다. 이 점수는 블로거가 작성한 모든 페이지들의 수와 블로거가 작성한 전체 페이지 당 한 페이지의 트랙백 가치를 기반으로 계산한다. 이러한 생각을 실현하기 위해 블로거 활동점수(BAS)를 다음과 같이 정의한다.

$$BAS(b,t) = N^t_e(b) \times \frac{N^{tr}_e(b)}{N_e(b)} \quad (7)$$

where,

$N^t_e(b)$: 블로거 b가 태그 t에 대해 작성한 모든 페이지의 수,

$N^{tr}_e(b)$: 블로거 b가 작성한 페이지 중 트랙백연결을 가진 페이지의 총 수,

$N_e(b)$: 블로거 b에 의해 작성된 모든 페이지의 수.

식 (7)에서, 첫 번째 요소는 블로거 행동의 양적요소를 나타내고 마지막 요소는 블로거 행동의 중요도를 나타낸다. 블로거 행동의 중요도는 트랙백연결을 기반으로 구해진다. $N^{tr}_e(b)/N_e(b)$은 블로거가 작성한 모든 페이지의 수와 그 페이지들 중 트랙백을 가진 페이지의 수의 비율이다. 이것은 블로거가 작성한 하나의 페이지가 얼마만큼의 트랙백연결을 가지는지 나타낸다. 즉 블로거의 한 번의 행동이 얼마만큼의 트랙백연결기여도를 가지는지 평가하는 것이다. 따라서 식 (7)은 블로거 b가 특정주제 t에 대하여 작성한 페이지 수(행동의 양)와 작성 페이지들의 트랙백 연결성(행동의 중요도)으로 블로거의 행동성을 평가한다.

4.2 트랙백 점수(Trackback Score)

이 점수는 특정한 태그를 가진 페이지의 트랙백 연결성을 나타낸다. 1장에서 언급하였듯이, 트랙백 연결성은 트랙백연결의 수와 트랙백연결을 생성한 블로거의 명성을 기반으로 평가한다. 따라서 태그 t의 관점에서 페이지 e_i의 트랙백 점수는 다음과 같이 정의한다.:

$$TBS(e,t) = \sum_{i=1}^{K} BS(b_i, t) \quad (8)$$

where,

b_i: i번째 트랙백연결을 생성한 블로거 b,

$BS(b_i, t)$: i번째 트랙백연결을 생성한 블로거 b의 명성점수.

4.1장에서 정의하였듯이 $BS(b_i, t)$는 태그 t에 대한 블로거 b_i의 점수이다. 페이지의 트랙백연결 점수를 평가하기 위해 페이지의 트랙백연결 수와 트랙백연결을 생성한 블로거의 점수를 고려한다. 따라서 만약 더 높은 점수를 가진 블로거가 작성한 트랙백연결이 더 많으면 TBS는 더욱 높아질 것이다.

4.3 댓글 점수(Comment Score)

페이지 e의 댓글 점수는 다음과 같이 정의된다.

$$CS(e,t) = (\alpha_t^b + \beta)NC(e) \quad (9)$$

where,

$\alpha_t^b = \frac{N_{tr}(b,t)}{N_{cm}(b,t)}$

b: 페이지 e를 작성한 블로거,

$N_{tr}(b, t)$: 블로거 b가 태그 t에 대해 작성한 모든 페이지들의 트랙백연결 수,

$N_{cm}(b, t)$: 블로거 b가 태그 t에 대해 작성한 모든 페이지들의 댓글 수,

β: 0.001 (트랙백연결이 없는 경우를 위해),

$NC(e)$: 페이지 e의 댓글 수

$a^b{}_t$는 댓글의 가중치이다. 댓글은 트랙백에 비해 작성의 용이성이 더 좋으므로 트랙백의 가치보다 더 작다. 예를 들면, 사용자가 페이지를 본 후 약간의 관심이나 만족만이 있더라도 쉽게 댓글을 달지만 트랙백연결은 더 높은 관심과 만족감을 가질 때 작성한다. 왜냐하면 트랙백연결은 자신의 블로그와 연결을 해야 하므로 로그인 과정이 필요하여 댓글보다 작성과정이 더 복잡하며 자신의 블로거 방문자에게도 정보를 공유하고자 하는 의도가 있기 때문이다. 이로 인해 일반적으로 댓글의 수가 트랙백의 수에 비해 수십 배 많은 수로 존재한다. 따라서 댓글의 가치가 트랙백의 가치보다 더 작다고 볼 수 있으므로 $a^b{}_t$는 댓글의 영향력을 약화시킨다.

4.4 각 요소들 간의 상관관계

지금까지 살펴본 것처럼 제안하는 알고리즘은 블로그 영역의 각 요소들, 블로거, 트랙백연결, 댓글, 그리고 태그를 기반으로 블로거 점수(BS), 트랙백 점수(TBS), 댓글 점수(CS)를 계산하여 페이지를 평가한다. 하지만 이 세 가지 점수는 상호의존적인 관계를 가지며 페이지 평가 시 재귀적으로 계산된다. 먼저 블로거 점수(BS)의 계산식을 모두 전개해 보면 다음과 같다.

$$BS(b,t) = sigmoid\left(\frac{\sum_{e \in E_t^b}(1 + TBS(e,t) + CS(e,t))}{|E_t^b|}\right) + sigmoid\left(N_e{}^t(b) \times \frac{N_e{}^{tr}(b)}{N_e(b)}\right) \quad (10)$$

where, sigmoid(a) = 1 / (1 + e^{-a})

식 (10)에서 앞항의 분자에 트랙백 점수(TBS)와 댓글 점수(CS)가 포함되어 있는 것을 알 수 있다. 다음으로 트랙백 점수(TBS)를 구하는 식 (8)은 b_i(i번째 트랙백연결을 생성한 블로거 b)의 블로거 점수(BS)를 포함하고 있다. 댓글 점수를 구하는 식 (9)에서도 직접적인 수식을 포함하고 있는 것은 아니지만 트랙백연결의 개수를 기반으로 점수계산을 하고 있다. 따라서 제안하는 알고리즘은 블로그의 속성에 따라 크게 3개의 점수를 계산하여 페이지 점수를 계산하지만 각각은 서로 상호연관성을 가지고 재귀적으로 계산되어 지고 있으며 모든 계산은 트랙백연결에 영향을 받으므로 사용자 평가를 내포하는 연결성 분석에 기반을 두고 있다.

5. 실 험

5.1 실험 설정

블로그-랭크 알고리즘의 궁극적인 목적은 질의어와 반환되는 검색결과의 연관성을 기반으로 검색결과랭크의 효율성을 개선하는 것이다. 이러한 효율성 개선을 증명하기 위해, 본 논문에서는 제안한 알고리즘을 적용한 블로그 검색시스템을 구현하였다. 블로그 검색시스템의 구현을 위해 티스토리 도메인에서 200개의 블로그, 62906개의 페이지를 수집하였다. 또한 구현한 검색시스템의 검색효율성과 기존의 블로그 검색시스템의 검색효율성을 비교하였다. 신뢰할 수 있는 기존의 블로그 사이트들은 모두 영리업체들이므로 자사의 검색시스템에 대한 모든 기술적 정보를 비공개로 하고 있다. 이에 기존의 블로그 사이트들이 제안한 알고리즘과 유사한 검색기술(블로그의 각 특징적 속성활용)을 사용한다고 보증할 수는 없지만 국내에서 가장 인지도가 높고 사용자의 활동이 활발한 티스토리를 실험 도메인으로 선정하였다.

더 높은 검색효율성은 질의어와 연관성이 더 많은 페이지가 그렇지 않은 페이지보다 더 높게 랭크되는 것이며 결국 질의어와 검색결과 페이지간의 연관성 평가가 필요하다. 3장에서 언급한 것처럼 블로거는 자신의 페이지 내용을 잘 나타내기 위해 신중히 태그를 선택한다. 그러므로 본 논문에서는 질의어와 관련 있는 태그를 더 많이 포함하는 페이지가 질의어와 더 많은 연관성을 가진다고 가정한다. 즉, 검색결과페이지가 질의어와 관련이 있는 태그들을 더 많이 포함하고 있다면, 그 페이지는 질의어와 더 연관성이 높다고 보는 것이다.

질의어와 관련이 있는 태그들은 동시출현빈도수 방법으로 각각의 블로그 도메인에서 추출하였고 이를 연관 태그라 정의한다. 동시출현빈도수 방법은 다음과 같다. 만약, 질의어와 동일한 태그를 포함하는 페이지가 있다면, 그 페이지 내의 모든 태그들은 질의어와 1의 연관성을 가진다. 이렇게 1의 연관성을 가진 태그들이 다른 페이지에서 또 다시 동시에 존재할 때 마다 연관성 값은 1씩 증가하게 된다. 이러한 과정이 반복되어 연관성 값이 10을 넘는 태그들(즉, 질의어와 동일한 태그와 같은 페이지에 10번 이상 존재했던 태그들)을 선택하는 방법이다. 또한 이러한 태그들을 티스토리, 이글루스[13], 블로그코리아[14] 도메인의 페이지에서 추출한다. 여기서 티스토리는 블로그-랭크 알고리즘의 랭크계산을 위해 수집되었던 도메인이고 이글루스와 블로그코리아는 그렇지 않은 도메인이다. 실험의 공정성을 위해 연관 태그 추출에 활용한 이글루스 페이지들은 랭크계산을 위해 적용되었던 페이지들을 제외하였다.

검색결과의 효율성을 평가하기 위해 페이지에 포함된 질의어와 연관된 태그의 수를 측정하고 그 수를 K순위까지의 Normalized Discounted Cumulative Gain(NDCG at K)에 적용하였다. NDCG at K는 검색결과의 랭크 정확성을 측정하는 정보검색의 평가 metric중 하나이다[15]. 주어진 질의어 q가 있을 때 랭크된 검색결과는 다음의 식에 의해 K순위별로 NDCG가 계산된다.:

$$NDCG_q^K = M_q \sum_{j=1}^{K} (2^{r(j)} - 1) / \log(1+j)$$

NDCG at K는 검색결과의 순위 1에서 K까지의 gain의 합으로 계산된다. $r(j)$는 순위 j에서의 보상을 나타내는 함수이다. 본 논문의 실험에서 $r(j)$는 j번째 페이지 e_j의 $RV(ej, q)$를 기반으로 계산된다.: $r(j) = log(RV(e_j, q) + 1)$. $RV(e_j, q)$는 페이지 e_j에서 질의어 q와 연관 태그들의 개수다. M_q는 정규화 상수이고 이로 인해 NDCG의 최댓값은 1이 된다.

실험의 현실성을 위하여, 질의어 풀에서 임의로 선택된 질의어를 이용하여 실험하였다. 총 20개의 실험질의어는 10개의 이슈질의어와 10개의 임의질의어로 구성한다. 이슈질의어는 최근에 사회적으로 이슈가 되는 사건이나 인물들과 관련된 단어들을 질의어 풀에서 선택하였고, 예를 들면, 이명박, 광우병, 올림픽 등등, 임의질의어는 질의어 풀에서 임의로 선택하였다. 질의어 풀은 다음의 과정으로 구성한다.

(1) 초기 단어를 가진 집합을 하나의 질의어-집합으로 둔다. 초기 단어는 임의로 선택한다.
(2) 질의어-집합에서 단어를 하나 임의로 선택한다.
(3) 선택된 단어와 관련된 단어를 블로그 영역에서 검색하고 검색된 단어를 질의어-집합에 추가한다. 관련된 단어는 선택된 단어와 동일한 태그를 가진 페이지의 모든 태그들이다.
(4) 질의어-집합의 크기가 100을 넘으면 중단한다.
(5) 이전에 선택되지 않은 단어를 하나 선택하고 3번 과정으로 돌아간다.

완성된 질의어-집합을 질의어 풀로 이용한다.

5.2 실험 결과

제안한 알고리즘을 적용한 블로그 검색시스템의 검색 효율성을 평가하기 위해 4장에서 정의한 설정을 따라 다음과 같이 실험하였다.

(1) 실험 질의어들을 구성한다. 실험 질의어들은 10개의 이슈질의어와 10개의 임의질의어로 구성된다.
(2) 실험 질의어들을 블로그-랭크 알고리즘을 적용한 검색시스템과 기존의 티스토리 블로그 검색 시스템에 각각 적용하고 검색결과를 받는다.
(3) NDCG at K 평가 metric과 3개 도메인에서 추출한 연관 태그들을 이용하여 두 시스템의 검색결과를 비교 분석한다.

페이지가 질의어와 관련 있는 연관 태그들을 더 많이 포함하고 있다면 그 페이지는 질의어와 관련 있는 정보를 포함하고 있을 가능성이 높다. 따라서 포함된 연관 태그의 수는 질의어와 페이지가 포함한 정보간의 관련성 유무를 판단할 수 있는 하나의 판단지표로 활용될 수 있다. 하지만, 페이지가 포함한 정보와 질의어와의 관련성을 포함하는 연관 태그의 개수만으로 평가하는 것, 즉 평가지표로 활용하는 것은 블로거, 트랙백 연결, 댓글과 같은 블로그의 여러 가지 속성들을 고려한 평가보다 더 신뢰도가 높은 검색결과를 기대하기는 힘들 것이다. 왜냐하면 질의어와 관련 있는 페이지가 연관 태그를 가질 수는 있지만, 연관 태그를 가진 모든 페이지가 질의어와 관련 있는 페이지로 볼 수는 없기 때문이다.

먼저, 티스토리 도메인에서 추출한 연관 태그들을 적용하여 실험하였다. 그림 2는 이슈질의어에서의 검색성능 개선을 보여준다.

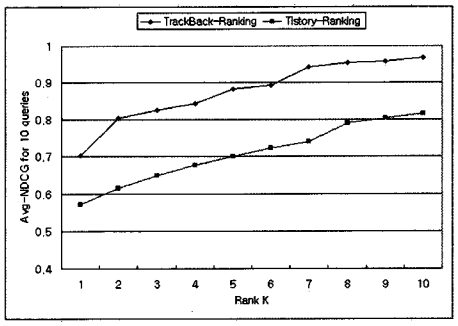


그림 2 이슈질의어와 티스토리의 연관 태그 기반 Avg-NDCG at K

그림에서 보인 것처럼, 상위 10위까지의 페이지에서 전체적으로 지속적인 성능개선이 나타났다. 그림의 NDCG 값은 10개 이슈질의어들의 평균값을 나타낸 것이다. 랭크 2(K=2)에서, 블로그-랭크의 NDCG 값은 0.804이고 티스토리-랭크의 값은 0.615이다. 랭크 1에서 2사이의 그래프 기울기를 살펴보면 블로그-랭크의 기울기가 다른 구간에 비해서 급격하게 변화한다. 이것은 랭크 2의 점수가 다른 랭크들에 비해 더 좋은 점수를 받았다는 것을 의미한다. 왜냐하면 NDCG at k값은 k랭크까지의 누적점수를 나타내기 때문이다. 티스토리-랭크의 기울기 변화를 보면 랭크 7과 8사이에서 가장 큰 기울기 변화를 보인다. 전체적인 기울기 변화를 보면 블로그-랭크의 상위랭크들이 하위랭크들에 비해 더 좋은 점수를 받았다는 것을 알 수 있는 반면 티스토리-랭크는 하위랭크가 더 좋은 점수를 받았다는 것을 알 수 있다.

그림 3은 임의질의어에서의 검색성능 개선을 보여준다.

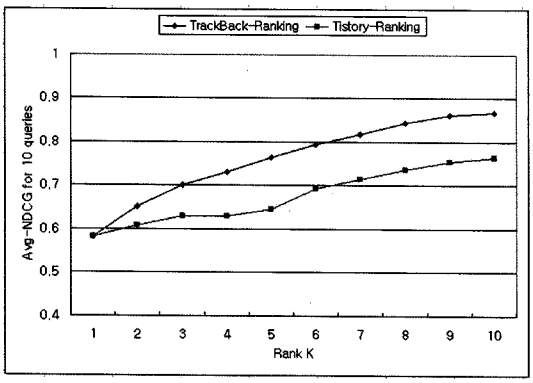


그림 3 임의질의어와 티스토리의 연관 태그 기반 Avg-NDCG at K

임의질의어에서도 상위 10위 페이지에서 전체적으로 지속적인 성능개선을 나타냈다. 가장 큰 성능개선을 보인 곳은 랭크 5에서이며, 블로그-랭크 0.764인데 반해 티스토리-랭크는 0.645이다. 전체적인 기울기변화를 살펴보면 블로그-랭크가 상위랭크에서 더 높은 점수를 받았다는 것을 알 수 있으며 티스토리-랭크는 중위랭크에서 높은 점수를 받은 것을 알 수 있다. 이와 같은 결과는 블로그-랭크 알고리즘이 기존 티스토리의 랭크알고리즘 보다 검색성능이 더욱 우수하다는 것을 보여준다. 또한 이슈질의어에서의 성능개선이 임의질의어에서의 성능개선보다 더욱 좋은 것으로 미루어 블로그-랭크 알고리즘이 최근에 이슈가 되는 페이지를 더 잘 반영한다는 것을 보여준다. 또한, 우리는 랭크계산에 적용된 도메인(티스토리)이 아닌 다른 두 도메인, 이글루스와 블로그코리아에서 추출한 연관 태그들도 활용하여 실험하였다. 그림 4, 5는 이글루스의 연관 태그를 적용한 이슈질의어와 임의질의어의 성능개선이고 그림 6, 7은 블로그코리아의 연관 태그를 적용한 이슈질의어와 임의질의어의 성능개선이다.

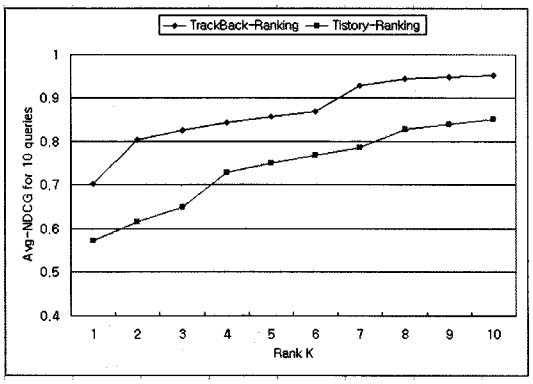


그림 4 이슈질의어와 이글루스의 연관 태그 기반 Avg-NDCG at K

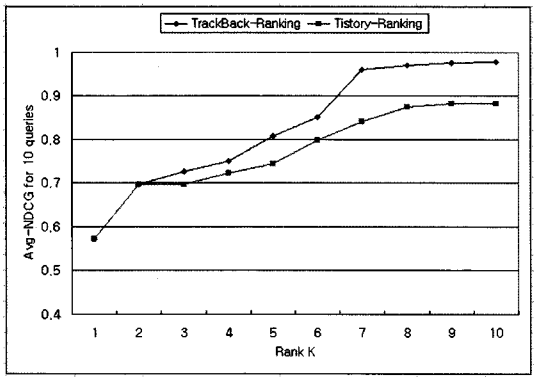


그림 5 임의질의어와 이글루스의 연관 태그 기반의 Avg-NDCG at K

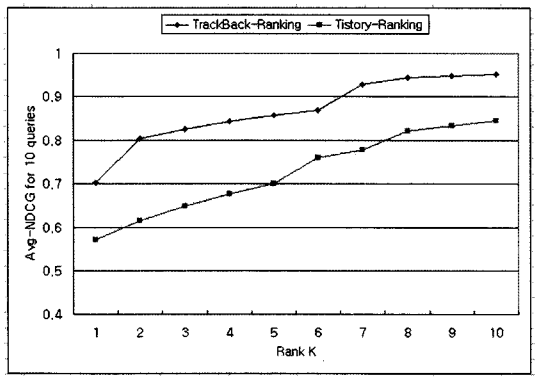


그림 6 이슈질의어와 블로그코리아의 연관태그 기반의 Avg-NDCG at K

이글루스와 블로그코리아의 연관태그를 적용한 실험

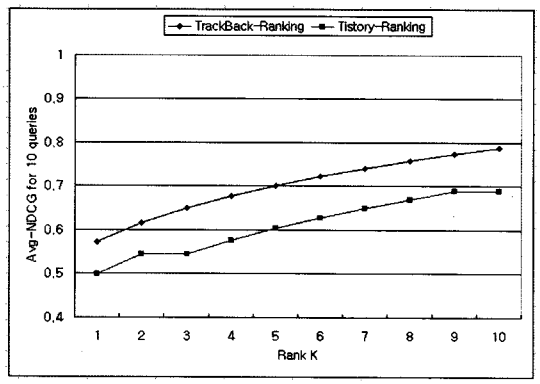


그림 7 임의질의어와 블로그코리아의 연관태그 기반 Avg-NDCG at K

표 1 블로그-랭크 알고리즘의 검색성능 개선

	연관태그들의 추출도메인		
	티스토리	이글루스	블로그코리아
이슈질의어	0.168 (24%)	0.129 (18%)	0.142 (20%)
임의질의어	0.085 (12%)	0.057 (7%)	0.090 (15%)

에서도 제안한 알고리즘의 성능이 더 우수한 것으로 나타났다. 이로써 블로그-랭크 알고리즘을 적용한 검색시스템이 기존의 티스토리 검색시스템의 검색성능보다 우수하다는 것을 보였으며 지금까지의 결과를 표 1에 정리하였다. 각 값은 블로그-랭크 알고리즘을 적용한 검색시스템과 티스토리 검색시스템의 Avg-NDCG값 차이이며 괄호안의 값은 제안한 알고리즘을 적용한 검색시스템의 성능개선 율이다.

지금까지 전체 실험질의어를 적용한 결과를 보였다. 실험결과에서 블로그-랭크 알고리즘은 기존의 검색시스템보다 전체 실험질의어에 대해 평균 16%의 성능개선을 보였으며 이슈질의어에 대한 성능개선이 평균 20%로 임의질의어의 평균 11% 보다 2배정도 높은 성능개선을 보였다.

6. 결론 및 향후연구

전통적인 웹 페이지와 다르게 블로그 페이지는 트랙백연결, 태그, 댓글과 같은 구조적인 특징을 가지고 있다. 그러므로 이러한 특징들을 고려하는 랭크 알고리즘이 기존의 웹 페이지 랭크 알고리즘보다 검색성능이 더 우수하다.

본 논문에서는 이러한 블로그의 구조적 특성들을 고려하여 더욱 효율적인 검색을 위한 블로그-랭크 알고리즘을 제안한다. 또한 알고리즘의 검색성능 향상을 평가하기 위해 블로그 검색시스템을 구현하여 기존의 시스템과 비교평가 하였다. 제안하는 알고리즘은 블로그 페이지를 크게 3가지의 요소로 분석하였다.: 블로거의 명성, 트랙백 연결성, 사용자의 반응성. 이러한 요소 중, 트랙백 연결성과 사용자 반응성은 단지 블로그의 구조적인 요소만을 평가하지만 암묵적인 내용평가를 포함하여 더 우수한 검색성능을 냈다. 이것은 컴퓨팅 성능의 관점에서 중요한 특징 중의 하나이다. 왜냐하면 데이터의 내용을 직접분석 하는 것은 비싼 컴퓨팅 비용이 요구되기 때문이다. 또한 제안한 알고리즘은 블로거의 명성을 평가 및 관리하여 초기에 연결성을 가지지는 않지만 유용한 내용을 가진 페이지를 저평가 하지 않는다.

앞으로의 연구는 스팸 트랙백과 댓글을 정제할 수 있는 기술을 재발 및 적용하고 좀 더 범용적인 크롤러를 개발하여 실용적으로 활용 가능한 블로그 전용검색엔진에 관한 연구를 진행할 것이다.

참 고 문 헌

[1] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," In *Proceedings of 7th International World Wide Web Conference, Computer Networks and ISDN Systems*, vol.30, no.1-7, pp.107-117, Apr., 1998.

[2] J. M. Kleinberg, "Authoritative sources in hyperlinked environment," *Journal of the ACM*, vol.46, no.5. pp.604-632, Sep., 1999.

[3] C. Marlow, "Audience, structure and authority in the weblog community," In *International Communication Association Conference*, NewOrleans, LA, 2004.

[4] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the Spread of Influence on the Blogosphere," Technical report, University of Maryland, Baltimore County, May., 2006.

[5] http://www.tistory.com/

[6] E. Adar., L. Zhang., L. Adamic., and R. Lukose., "Implicit Structure and the Dynamics of Blogspace," *Workshop on the Weblogging Ecosystem: Aggregation*, Analysis and Dynamics, 2004.

[7] K. Fujimura., T. Inoue., and M. Sugisaki., (2005). "TheEigenRumor Algorithm for Ranking Weblogs," 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2005.

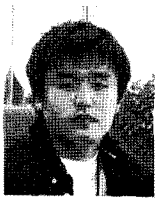
[8] G. Mishne, "Multiple Ranking Strategies for Opinion Retrieval in Blogs," In *Proceedings of TREC 2006*, 2006.

[9] G. Mishne, "Using Blog Properties to Improve Retrieval," In *Proceedings of the ICWSM*, 2007.

[10] K. Liu, G. Qiu, J. Bu, and C. Chen, "Ranking Using Multi-features in Blog Search," In *Advances in Multimedia Information Processing - PCM 2007*, Lecture Notes in Computer Science, vol.4810/2007, pp.714-723, 2007.

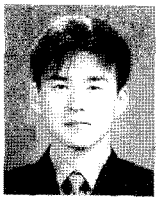
[11] A. Java, P. Kolari, T. Finin, A. Joshi, and J.

Martineau, "The BlogVox Opinion Retrieval System," In *Proceedings of TREC 2006*, 2006.
[12] P. Kolari, A. Java, T. Finin, J. Mayfield, A. Joshi, and J. Martineau, "Blog Track Open Task: Spam Blog Classification," *Technical report, September 2006*. TREC 2006 Blog Track.
[13] http://www.egloos.com/
[14] http://www.blogkorea.net/
[15] K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," In *Proceedings of the ACM Conference on Research and Development on Information Retrieval (SIGIR)*, pp.41-48, 2000.



김 정 훈

2006년 동아대학교 전기전자컴퓨터공학부 학사. 2009년 성균관대학교 전자전기컴퓨터공학과 석사. 2009년~현재 삼성전자 DMC연구소 연구원. 관심분야는 Semantic 검색, 지능시스템, 컴파일러



윤 태 복

2001년 공주대학교 전자계산학과 학사 2005년 성균관대학교 컴퓨터공학과 석사 2005년~현재 성균관대학교 컴퓨터공학과 박사과정. 2008년~현재 성결대학교 멀티미디어학부 외래교수. 관심분야는 사용자 및 상호작용 모델링, 게임 인공지능

이 지 형

1993년 한국과학기술원 전산학과 학사 1995년 한국과학기술원 전산학과 석사 1999년 한국과학기술원 전산학과 박사 2002년~현재 성균관대학교 정보통신공학부 부교수. 관심분야는 지능시스템, 기계학습, 온톨로지