韓國컴퓨터情報學會 論文誌
第15卷 第12號, 2010. 12.

2010-15-12-3-6

TAG 정보를 활용한 기업검색의 적합성 향상 기법에 관한 연구

손 태 식*, 박 병 섭**, 최 효 현***

A Study on the Relevance Improvement of Enterprise Search using Tag Information

Taeshik Shon*, Byoungseob Park**, Hyo Hyun Choi***

요 약

기업에서 업무 시스템들을 활용하여 업무를 진행하다 보면 기하급수적으로 증가하는 정보를 얼마나 신속하고 정확하게 사용자에게 제공할 수 있는가 하는 것이 기업 경쟁력의 중요한 요소이다. 검색 적합성 향상을 통한 양질의 검색 결과 제공은 기업 경쟁력의 중요한 요소가 되었으며, 가치 있고 효율적인 검색 서비스 제공을 위해 검색엔진에서 제공하는 단순한 검색 서비스 이상을 제공하는 것이 필요하다. 본 논문에서는 검색 과정에서 Tag 정보와 그 가중치 값을 활용하여 검색 적합성을 향상 시키는 방안에 대해서 연구함으로써 검색엔진에서 제공하는 검색 적합성의 한계를 극복하는 방안을 제안한다. 또한, 제안된 방법에 대한 검색 성능을 비교하기 위해서 제안 기법을 기존 웹 검색 서비스에서 제공하는 검색결과와의 적합성 평가 및 연관 검색어와 비교함으로써 우수성을 검증하였다.

Abstract

In this paper, how fast and accurate the companies provides exponentially increasing information to the users is the most important in the corporate competitiveness. The enhancement of the retrieval relevance became the important element in enhancing company competitiveness and it is required to provide the services that are beyond simple retrieval service for good quality search service. This paper proposes the effective scheme that enhances retrieval relevance by utilizing registered tag information. By proposed scheme, we can overcome the limitations of retrieval relevance that usual search engines provide. And we compare the proposed scheme with existing web retrieval service on retrieval relevance evaluation and related search keyword.

▸ Keyword : 검색 적합성(Retrieval Relevance), 기업 검색(Enterprise Search), 태그(Tag), 검색 엔진(Search Engine), 웹 검색(Web search)

• 제1저자 : 손태식 교신저자 : 최효현
• 투고일 : 2010. 08. 03, 심사일 : 2010. 09. 28, 게재확정일 : 2010. 10. 07.
* 삼성전자 Digital Media & Communications 연구소 책임연구원
** 인하공업전문대학 컴퓨터시스템과 부교수 *** 인하공업전문대학 컴퓨터정보과 조교수

Ⅰ. 서 론

가트너 보고서에 의하면 기업 검색(Enterprise Search)의 범위가 내부 정보에 대한 통합 검색에서 내·외부 정보의 통합 검색으로 그 범위가 확대되고 있으며, 정보 접근 방식에 있어서도 접근의 효율성을 높일 수 있도록 멀티미디어 검색, 지능적 검색 기능 등을 추가하는 방식으로 발전하고 있다[1]. 이러한 경향은 인터넷을 통한 지식정보의 증가로 사용자가 필요로 하는 정보를 얼마나 빠르고 정확하게 찾는가 하는 것이 검색 서비스의 중요한 화두가 되고 있음을 반증하는 결과로서, 기업 사용자에게 필요한 검색결과를 적시에 제공함으로써 사용자에게 정보를 가치 있는 지식으로 재생산하도록 도와준다. 일반적으로 검색은 개인화된 웹 검색과 특화된 사용자 계층을 위한 기업 검색으로 분류 될 수 있는데, 이러한 웹 검색과 기업 검색은 검색결과의 활용 측면에서 많은 차이를 보인다. 먼저, 웹 검색은 사용자들이 가치 있는 지식으로 사용하기 위한 초기 자료로 주로 활용하지만, 기업 검색은 검색된 특정 분야의 기술, 동향, 최신 정보를 바로 업무에 사용하기 위해서 주로 활용한다. 기업 검색은 검색결과에서 얻어진 내용을 업무 생산성 향상을 위해서 활용하므로 사용자가 원하는 정보를 정확하고, 신속하게 찾아주는 것이 기업 검색에 있어서 중요한 역할이다. 하지만 기업에서 활용하고 있는 검색엔진이 좋은 성능을 가지고 있다고 할지라도 일반적인 검색 랭킹 알고리즘에서 제공하는 검색 적합성(relevance) 이상의 결과를 사용자에게 제공할 수 없기 때문에 실제로 사용자가 필요로 하는 자료가 하위에 랭크되어 적시에 활용하지 못하거나 해당 자료를 찾기 위해서 많은 노력을 필요로 한다. 결국 기업이 보유하고 있는 그룹웨어, 지식 관리 시스템, 문서 관리 시스템 등에 대한 통합 검색 외에 해당 기업에 필요한 유관 기관 홈페이지, 국내·외 보고서, 연구 자료, 학술 정보 등 외부 전문 지식 정보를 주기적으로 모니터링하고 수집하여 업무에 활용할 수 있도록 하는 내·외부 통합 검색이 기업 검색의 방향이 될 것으로 예측되고 있는 것이다. 이러한 예측이 의미하는 바는 사용자별로 맞춤 검색 결과를 제시할 수 있도록 하는 개인화 검색과 외부 정보의 지속적인 수집을 가능하게 하는 외부 정보 수집 및 분석이 중요한 기능이 될 것이라는 것이다[2-3].

본 논문에서는 위와 같은 기업 검색의 방향과 더불어 실제 기업 검색에서의 검색결과 적합성 향상을 위해서 TAG 정보를 활용함으로 기존 검색엔진의 랭킹 알고리즘의 한계를 극복하고 검색 적합성 향상 및 검색어와 관련된 다양한 부가 서비스를 제공하여 사용자가 필요한 정보를 효과적으로 활용하도록 한다. Tag는 콘텐츠에 대해서 사용자들이 직접 의미 있는 키워드들을 꼬리표처럼 등록하는 것으로 트위터의 Hash Tag와 페이스북의 Photo Tag 같은 것들이 인기를 끌면서 Tag에 대한 관심이 더 높아지고 있다. Tag는 기계적으로 추출된 연관 키워드와 비교하여 명확하게 의미를 전달할 수 있기 때문에 Tag와 관련한 연관 정보를 분석함으로써 검색결과의 정확도를 향상 시킬 수 있다[4], 본 논문에서는 Tag 정보를 활용한 검색 적합성 향상 평가와 연관 검색어에 대한 성능 평가를 통하여 제안 방안을 검증한다. 이러한 제안 기법은 결과적으로 Tag를 활용한 검색 적합성 향상은 사용자가 필요로 하는 정보를 적시에 제공함으로써 업무 생산성 및 효율성을 증대하는데 기여 할 수 있다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2장에서는 기존에 알려진 기업 검색 기법 및 상용 기업 검색 엔진에 대해서 서술하며, 3장에서는 Tag기반 제안 기법을 설명한다. 4장에서는 Tag 정보를 활용한 검색 적합성 및 연관 검색어에 대한 평가 실험을 통한 제안 기법을 검증하고, 5장에서 본 논문의 결론 및 향후 연구방향에 대해서 서술한다.

Ⅱ. 관련 연구

인터넷으로부터의 최적의 적합한 정보를 검색하는 것은 많은 시간과 노력이 필요한 작업이며, 이때. 정확한 정보를 웹 검색만을 통해서 얻는 것은 힘든 작업일 수 있다.

하지만, 웹 검색은 대단위의 후보 페이지들로부터 다양한 방법을 통해 얻기 때문에 [5], 오늘날의의 검색 엔진은 인터넷으로부터 정보를 검색하고 검색하기 위한 가장 폭넓게 사용한 도구이다. 현재 다용도로 사용되고 있는 검색 엔진은 사용자로부터 검색에 필요한 정보를 충분히 얻지 못하는 경우 종종 부족한/부적절한 결과를 제공하기도 한다. 즉, 사용자는 몇몇 키워드를 검색 엔진에 제공하고, 검색엔진은 그 키워드를 포함한 웹 페이지들의 링크를 제공하는 것이 일반적인 방식이기 때문이다 [6]. 보편적으로 웹 검색 엔진의 사용자들은 짧은 질문을 이용하는 경향이 있다는 것은 이미 잘 알려진 사실이며, 실제로 사용자들의 짧은 질문은 단지 하나 또는 두 단어로 구성된다[7-8]. 그러나 짧은 질문은 사실상 대부분 한정적이거나 특정적인 문구를 포함하지 못하고 있는 것이 일반적이기 때문에 유사성 기반 방법에 의해 검색된 결과는 사용자에게 매우 낮은 수준의 검색 결과를 종종 제공 할 수 있다. 이러한 문제점은 사용자들에게 품질 정보를 검색하고 검색한 데 상당히 많은 시간을 쓰도록 유도한다. 이러한 문제를

고려함에 있어, 개인화된 웹 검색 기법의 적용이 다양한 관점에서 거론되고 있다 [9-10]. 하지만, 개인화된 검색 서비스를 제공하는 것은 쉽지 않은 문제로서 개별 사용자들의 관심 분야를 파악하거나 검색어 로그 기록을 분석하여 사용자가 필요로 하는 검색정보를 개인화하여 제공하는 것은 아직까지는 해결할 많은 문제점을 가지고 있다. 이러한 문제점을 보완하기 위한 한 방법으로서 다양한 웹 에이전트들을 활용하여 검색하는 기법이 제안되고 개발되었다[11].

다음으로 현재 기업 검색에 활용되고 있는 몇몇 알려진 솔루션들의 특성을 살펴본다. 먼저 Inxight는 텍스트 분석을 위한 종합 솔루션으로서[12] 'Inxight Smart Discovery Server'라는 검증된 엔터프라이즈 검색 엔진과 개체명 추출기인 'ThingFinder'와 결합하여 검색 결과를 자동으로 분류해 준다. 이는 사용자가 검색 결과에 있는 사람, 회사, 장소, 개념 등의 정보를 기준으로 검색 결과를 걸러 낼 수 있게 하는 특화된 기능이다. 또한, 개인화 검색을 통해 새로 갱신된 개인 관심 정보를 자동으로 알려준다.

CISCO는 출판 콘텐츠에 대한 신속한 대응을 위해 검색과 자동화된 감정 분석을 결합한 'Second Opinion' 시스템을 도입하였다. 'Second Opinion'은 출판 미디어, 분석 미디어, 고객 만족에 대한 서베이, 블로그 등을 분석하여 종합적으로 보고하는 시스템이다. 현재 12,000여 개의 웹 출판 사이트와 Lexis-Nexis를 통한 인쇄 출판물로부터 13개의 CISCO 기술 분야, 36개의 경쟁사, 550개의 제품 라인, 45,000여 개 이상의 회사들에 대한 데이터를 수집한다. 내부적으로는 Lexalytics의 'Salience Server'를 기반으로 순수 통계 처리와 품사 태깅, 문장과 단락 구조 분석, 단어 의미와 어조(Tone)에 대한 통계 분석 등을 수행한다. 이 과정을 통해서 사람/장소/상표/회사 등의 개체명을 찾고 이에 연관된 어조를 할당하고 나아가 문서의 어조를 결정한다 [13].

마지막으로 'Search Formula-1 Enterprise Edition'은 대용량, 고속 처리를 강점으로 커스터마이징이 용이할 수 있도록 유연성과 확장성을 고려하여 개발된 검색 엔진이다. Search Formula-1은 초기 설계 시부터 대량의 데이터에 대한 신속한 처리, 안정적인 구동, 효율적인 리소스 사용까지 감안하여 개발된 검색 엔진으로 보다 효율적인 커스터마이징을 가능하게 한다. 최근에 검색 엔진에게 요구되고 있는 대용량 분산 처리, 다양한 플랫폼 및 DBMS 지원, 동적 색인, 검색 엔진 통합 관리 기능을 모두 포함한다[14].

하지만 아직 이렇게 알려진 기업 검색 솔루션일지라도 대부분 랭킹 기반 검색 기법을 사용하고 있으며, 그 외 최근 흐름인 멀티미디어, 내/외부 통합, 지능화 검색, 개인간 업무 특성 반영 등의 부수적인 기능 추가에만 초점을 맞추고 있다. 즉, 본 논문에서는 의미 있는 키워드인 Tag를 활용하여 검색 적합성을 높일 수 있는 방안을 제시하려 한다. 보다 자세한 제안 방안의 내용은 다음의 3장에서 서술된다.

III. Tag 정보 기반 제안 기법

일반적인 웹의 관련 검색 결과는 전체 구조 설계, 세부 내역과 배포 계획과 같은 3가지 방향으로부터 의미적 접근을 재고려하는 것으로 구축 될 수 있다. 검색 엔진은 일정한 기준에 따르면 목록을 순서화하기 위해 등급책정 알고리즘을 사용한다. 랭킹알고리즘은 색인 문서의 키워드를 기반으로 우선순위를 정의하기 때문에 검색어에 해당하는 키워드를 포함하지 않을 경우 검색결과에서 제외되거나 검색결과의 하위에 디스플레이 된다. 하지만, 검색 문서가 검색어를 포함하지 않는다고 하여 중요도가 떨어지거나 사용자가 필요로 하는 문서가 아니라고 볼 수는 없기 때문에 이 문제를 해결하는 것이 검색 적합성을 향상 시키는 중요한 요소이다. 따라서 본 논문에서는 Tag 정보를 활용하여 콘텐츠에 대해서 사용자가 직접의미 있는 키워드를 부여함으로써, 이에 따라 기존 랭킹 알고리즘을 통한 검색 적합성 방식의 문제점을 해결하고자 한다. 검색어를 포함하는 Tag의 콘텐츠는 기계적인 방식으로 추출된 검색 적합성에 비하여 높은 적합성을 가지고 있다고 볼 수 있으며, 함께 등록된 Tag 목록은 검색어에 대한 연관 검색어로 활용되어 적합성 향상을 위한 연계 정보로 활용하게 된다.

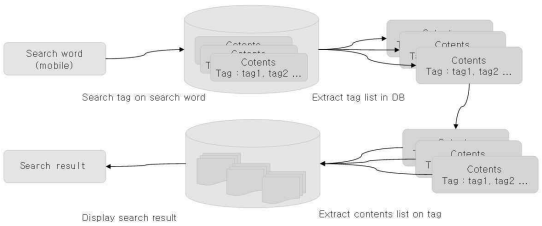

그림 1. Tag 정보를 사용한 제안 구조의 전체 구성
Figure 1. The Architecture on improvement of search relevance using Tag Information

따라서 본 장에서는 Tag를 활용한 검색 적합성 향상 및 연관된 부가 서비스 제공을 위한 구조를 제안하고, 연관 Tag 정보를 활용한 검색 서비스를 위한 알고리즘을 살펴본다. 다음의 그림 1은 제안된 구조의 전체 흐름도이며, 검색어에 대한 검색 적합성 계산을 위해서 6단계의 절차를 수행한다. 각 단계에 대한 세부 제안 내용은 각 세부 장에서 논의 된다.

1단계 : 입력한 검색어를 포함하는 콘텐츠 목록을 추출
2단계 : 검색된 콘텐츠들과 함께 포함된 연관 Tag 리스트 추출
3단계 : relation tag들을 포함하는 콘텐츠 목록 추출
4단계 : tag 정보를 활용한 검색 적합성 알고리즘 적용
5단계 : 검색어에 대한 검색 결과 표현 방식 UI
6단계 : 검색어에 대한 연관 검색어 표현 방식 UI

3.1 검색어를 포함하는 콘텐츠 리스트 추출

콘텐츠에 대해서 사용자는 N개의 Tag를 등록 할 수 있으며, Tag는 키워드 기반으로 콘텐츠의 의미를 정의하고 있다. 콘텐츠에 등록된 Tag는 등록자가 의미 있는 키워드를 부여한 것이므로 빈도수에 의해서 추출된 키워드에 비하여 더욱 활용성이 높으며, 기계적으로 추출된 키워드와 비교하여 Tag를 활용하는 것이 콘텐츠의 의미 부여에 효율적이다. 콘텐츠에 등록된 Tag들이 검색어를 포함하고 있으면 콘텐츠 목록을 추출하여 Contents_List에 저장한다.

$CLi = \{ Ci : T \in Ci \}$

(Ci은 검색어를 포함하는 Tag의 콘텐츠, T는 검색어와 일치하는 Tag)

그림 2. 콘텐츠에 등록된 Tag 정보 리스트
Figure 2. The registered Tag list on contents

3.2 콘텐츠에 함께 포함된 연관 Tag 추출

검색어에 의해서 추출된 콘텐츠들의 Tag들은 검색어와 연관된 Tag들을 함께 포함하고 있는데, 예를 들어 그림 2와 같이 노키아의 Touch UI에 대한 기술 관련 소식들을 등록할 때 사용자들은 iPhone, Apple, Nokia, Touch Screen과 같은 Tag들이 해당 콘텐츠를 대표하는 키워드로 생각하여 함께 등록한다. 사용자에 의해서 등록된 Tag들은 상호 의미 있는 연관성을 갖게 되므로 등록된 모든 Tag들을 Relation_Tag(rt)에 저장한다.

3.3 Relation Tag를 포함하는 콘텐츠 리스트 추출

검색어를 포함하는 콘텐츠의 Tag들은 연관된 Tag를 포함하고 있는데, 연관 Tag는 사용자가 찾고자 하는 콘텐츠를 확장하여 검색할 수 있도록 한다. 예를 들어 사용자는 Nokia라는 검색어로 검색하면 검색엔진은 Nokia란 색인어를 포함하는 검색결과를 디스플레이 하지만, 연관 Tag를 활용한 검색은 사용자가 입력한 iPhone, Apple, Touch Screen등을 포함하는 Tag의 콘텐츠를 검색결과에 디스플레이 한다. 연관 Tag들을 포함하고 있는 Tag의 콘텐츠 목록을 추출하여 Relation_Contents_List에 저장하며, 연관 Tag들의 연결 Depth에 따라 가중치를 부여한다.

$RCLi = \{ Ci : Tj \in Ci \}$

(Ci은 Relation Tag를 포함하는 Tag의 콘텐츠, Tj 는 Relation Tag를 포함하는 Tag 리스트)

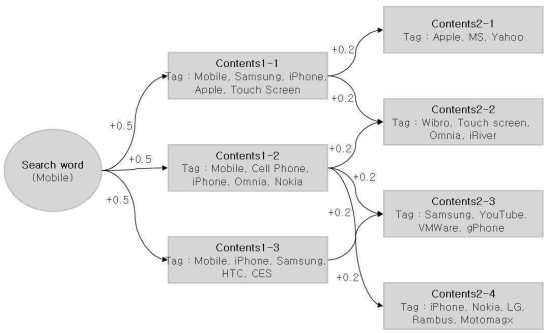

그림 3. 검색어에 대한 연관 Tag의 가중치 흐름도
Figure 3. Flowchart of Relation Tag according to Search words

3.4 Tag 정보를 활용한 검색 적합성 알고리즘

검색어에 대한 적합성 계산을 위해서 검색엔진들은 tf*idf 알고리즘을 기반으로 확장된 랭킹 알고리즘을 적용하고 있다. 하지만 기존 랭킹 알고리즘들은 검색어를 포함하는 색인어가 존재하지 않을 경우 검색어와 연관성을 가지는 콘텐츠지만 검색결과에 포함되지 않는데, 이 문제를 해결하기 위해서 Semantic 기반의 검색 기술을 연구하고 있지만 아직까지 해결할 많은 문제들 때문에 명확한 답을 제공하지 못하고 있다.

검색엔진에서 제공하는 적합성 결과와 본 논문에서 제안한 적합성 알고리즘을 결합하여 사용자에게 의미 있는 검색결과를 제공하고자 한다. 제안된 알고리즘의 상세내용은 다음의 그림 4와 같다.

```
WHILE (CL is not empty)
/* CL(Contents List) includes a identical tag with the search word */
    RelatedTagArray = Tags included in Contents List
    clScore = SE Score + weight 0.5
    /* SE Score is ranking score of Search Engine */
END WHILE

WHILE (D is not empty) /* D is all documents */
    IF Di includes RelatedTagArray
    RCL = Di    /* RCL(Related Contents List) is contents
                   including RelatedTagArray */
    END IF
END WHILE

WHILE (RCL is not empty)
     rclScore = SE Score + weight 0.2
END WHILE
```

그림 4. 본 논문에서 제안하는 검색 적합성 알고리즘
Figure 4. Proposed Relevance Search Algorithm

clScore는 검색엔진에서 계산된 적합성 점수와 검색어 Tag를 포함하는 콘텐츠에 가중치를 적용한 것이며, rclScore는 검색엔진에서 계산된 적합성 점수와 연관 Tag를 포함하는 콘텐츠의 가중치를 적용한 값이다. 계산된 clScore와 rclScore를 랭킹 알고리즘에 의해서 계산된 Score에 적용함으로써 검색결과에 가중치를 부여하여 적합성을 향상시킨다. 예를 들어 검색엔진의 랭킹알고리즘에 의해서 계산된 적합성 Score가 0.25일 경우 검색어와 동일한 Tag를 포함한 콘텐츠일 경우 가중치를 적용하면 Score가 0.75가 되고, 연관 Tag를 포함하는 콘텐츠일 경우에는 0.2의 가중치를 적용하면 0.45가 되어 검색결과의 우선순위가 조정되어 검색결과가 상위에 디스플레이 된다.

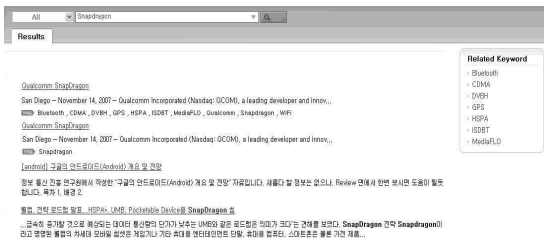

그림 5. 검색 결과 화면
Figure 5. The Screen of Search result

3.5 검색결과 및 연관 검색어 UI

검색어에 대한 검색결과는 제안된 방식의 검색결과를 웹 브라우저를 통해서 인터넷으로 서비스 한다. 사용자에게 제공되는 검색결과는 그림 5와 같이 콘텐츠의 제목, 등록자, 등록일, 요약글, 관련 Tag들로 구성되며, 정렬 타입에 따라 적합성, 최신등록일 순으로 디스플레이 한다. 또한 콘텐츠에 등록된 관련 Tag 정보를 활용한 연관 검색어 정보를 제공하는데, 기계적 학습에 의해서 추출된 Key Word 보다 뛰어나 성능을 보여준다. 상세한 성능비교는 4. 성능평가에서 설명하고, "touch screen"에 대한 연관 검색어 결과는 그림 6과 같다.

그림 6. Tag 정보를 활용한 연관 검색어 결과
Figure 6. The Relation word of search word using Tag Information

IV. 제안 기법의 성능 평가 및 분석

4.1 Tag 정보를 활용한 검색 적합성 향상 평가

본 장에서는 앞서 3장에서 제안된 Tag 정보를 활용한 검색 적합성 향상 알고리즘을 평가하기 위해서 제안 방안을 구현한 기업 검색용 프로토 타입 시스템을 활용하여 성능 평가 테스트를 진행했다. 제안 시스템이 적용된 프로토타입은 자바로 구현된 오픈소스 검색엔진인 Nutch를 활용하여 구현하였다[15]. Lucene이 Indexer와 Searcher로 구성되어 있고, Nutch는 Lucene에 없는 웹검색에 필요한 모든 기본요소를 전부 갖추어서 웹검색 용으로 확장 한 것으로서 본 논문의 프로토타입 시스템에서는 Nutch의 Indexer 및 Searcher 모듈에 추가로 Tager 모듈을 구현하였다. 제안된 알고리즘을 평가하기 위해서 검색 엔진에서 제공하는 일반적인 검색결과와 비교 검증 하였다.

본 논문의 제안 기법은 주로 Tager 모듈의 구현과 Tager 모듈을 통한 결과 분석이 주를 이루는데 Tager 모듈은 처음 사용자가 입력한 keyword들이 indexer 모듈에서 기존 indexing 정보와 비교되고 이후 비교 결과를 searcher 모듈

을 통해서 검색 결과를 표시하게 된다. 하지만 본 논문의 방식은 이 결과만을 보여주는 것이 아니라 Tager 모듈에서 기존 사용자들에 의해 입력된 keyword와 이에 연관된 tagging 정보를 별도 관련하여 단순한 indexing값과 이에 의한 검색 결과가 아닌 연관 검색 결과까지 모두 통합하여 제공해주는 것이다.

표 1. 제안 기법을 활용한 경우의 랭킹 변화
Table 1. Ranking change by proposed algorithm

Search word (result number)	Title of Results	Legacy (score/rank)	Proposed (score/rank)
Touch screen (476)	Multi-Touch Systems that I Have Known	0.45/21	0.95/5
	Prototype goes 'see through' with touch	0/0	0.5/21
	Toshiba, announce automobile HDDVD	0/0	0.5/22
LCD (2375)	China electronics industry R & D study	0.23/148	0.73/14
	GE Electronic business seeing in	0/0	0.5/32
	GE India-new manufacturing unit at	0/0	0.5/33
AJAX (369)	Asynchronous request about JavaScript and	0.11/243	0.61/42
	example of page call using ajax in jsp	0.28/142	0.78/16
	To learn Ajax	0.20/184	0.70/23

Table 1은 제안된 알고리즘을 적용했을 때 검색결과에 대한 랭킹의 변화를 나타낸다. 예를 들어, "Multi-Touch Systems that I Have Known"란 제목을 가진 콘텐츠는 검색 엔진에서 제공하는 Score가 0.45이지만 제안된 방식을 적용하면 0.95가 되어 검색결과의 상위에 위치하게 된다. 그리고 "Prototype goes 'see through' with touch" 제목의 콘텐츠의 검색결과는 검색 엔진에서는 제공하지 않지만, 제안된 방식을 적용하면 score가 0.5가 되어 검색 결과에서 디스플레이 되어 사용자에게 서비스된다.

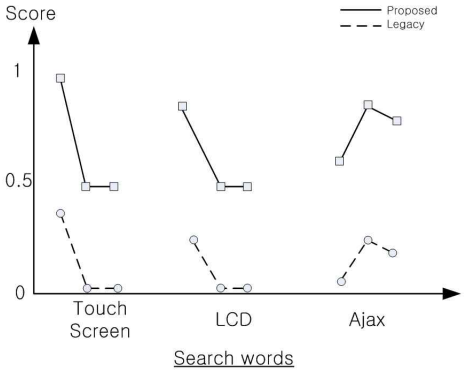

그림 7. 검색 단어에 따른 검색 결과의 스코어 비교
Figure 7. Comparison of Search words by Score

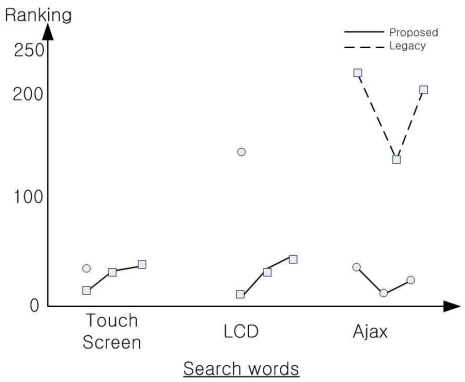

그림 8. 검색 단어에 따른 검색 결과의 랭킹 비교
Figure 8. Comparison of Search words by Ranking

따라서 그림 7과 같은 검색 단어별 스코어 비교 결과에 따르면 제안 기법은 "Touch Screen", "LCD", "Ajax"와 같은 주제의 검색에 있어서 기존 방법 대비 약 2배 이상의 높은 스코어를 보였으며, 그림 8의 분석과 같이 랭킹 분석 결과 적절한 검색 결과를 보이지 못해 랭킹 정보 자체가 부여되지 못하던 기존 방법과 달리 대부분 20위 이내의 상위에 위치하였음을 알 수 있었다.

4.2 연관 검색어에 대한 성능 평가

검색 서비스들은 사용자들이 입력한 검색어에 대한검색결과뿐 아니라 연관 검색어를 함께 제공하고 있다. 일반적인 검색서비스들의 연관 검색어는 수집된 콘텐츠를 분석하여 추출된 Key word의 빈도수를 기반으로 연관 검색어를 구성하지

만, 제안된 방식은 사용자가 입력한 Tag들을 기반으로 연관된 콘텐츠들에 포함된 Tags들의 목록을 빈도수로 계산하여 Score에 따라 디스플레이 한다. 본 장에서는 제안된 방식을 통해서 추출된 연관 검색어와 Google과 Ask.com에서 제공하는 연관 검색어의 성능을 평가하였다.

표 2. Tag 정보를 사용한 연관 검색어 검색
Table 2. Related Search word using Tag information

Search Word	Google	Ask.com	Proposed Method
iPhone	iPhone 3g	iPhone Accessories	Apple
	iPhone sdk	iPhone Release Date	3g iPhone
	Apple	iPhone iPhone Reviews	Smart phone
Apple	Apple store	Apple iPod	iPhone 3G
	Apple korea	Apple the Fruit	iTunes
	Apple ipod	Apple Tree	Droid Players
Solar Cell	Organic solar cell	Solar Cell Types	Organic solar cell
	Solar cell process	Solar Panels	Plastic solar cell
	Solar cell module	Solar Energy	DOE

표 2는 검색어에 대한 연관 검색어 성능을 비교 평가한 결과인데, 중요한 차이점은 Google과 Ask.com는 검색어를 포함한 형태의 연관 검색어를 제공하고 있으며, 제안된 방식은 검색어를 포함하지 않더라도 사용자가 입력한 Tag 정보를 추출하여 검색어와 연관된 키워드를 제공한다. 예를 들어, iPhone에 대한연관 검색어를 살펴보면 Google과 Ask.com은 iPhone을 포함한 연관된 검색어를 보여주지만, 제안된 방법은 연관 검색어로 Smart Phone를 제공함으로써 Mobile Phone 업체정보를 파악하거나 경쟁 업체의 Smart Phone 정보를 검색하도록 지원하는 등의 높은 효용성을 보여주었다.

V. 결 론

급변하는 컨버전스 환경에서 기업의 경쟁력을 강화시키는 수단의 하나로써 기업 내외부의 전사적 자원을 검색하여 효율적이고 가치있는 정보를 재생산 할 수 있는 검색 기술이 많은 각광을 받고 있다. 하지만, 기존 랭킹 기반의 검색에서의 의존성을 탈피하고 좀 더 의미있는 검색 데이터를 추출하기 위한 방안으로서 본 논문에서는 Tag 정보를 활용하는 방안을 제시하였다. 기존 랭킹 기반의 검색 기능을 확장하여 다양한 콘텐츠들의 특성을 반영할 수 있는 Tag 정보 및 적합성 결과에 따른 스코어 등을 활용한 제안 방안은 사용자에게 보다 높은 수준의 정확성을 제공하는 것은 물론이거니와 검색 후 2차 검색의 효율성을 가져다 줄 수 있는 솔루션 중에 하나가 될 수 있다. 제안 방안은 Tag 정보를 활용한 scoring, ranking 분석 및 연관 검색어 분석을 통해 기존 웹 검색 대비 두 배 이상의 높은 랭킹 및 스코어를 보여 줌을 알 수 있었다.

향후 연구에서는 입력된 Tag들의 연관 관계를 분석하여 Tag 기반의 Ontology 검색을 진행할 예정이다.

참고문헌

[1] 코리아와이즈넛, “Personal Workplace의 완성,” KM&EDM Korea Conference Fall, 2006.

[2] 정한민, 이승우, 성원경, “엔터프라이즈 검색 기술 동향”, 주간 기술 동향, 2006.

[3] G. Grefenstette. "Enterprise search trends and challenges." CHORUS Final Conference, 2009

[4] K. Bischoff, C.S Firan, W. Neidl, R. Paju, "Can all tags be used for search?," ACM conference on Information and knowledge management, pp. 193-202, 2008.

[5] J. Convey, “Online Information Retrieval Library Association Publishing," London, 1992.

[6] B. Jansen, A. Spink, J. Bateman, J. Saracevic, "Real life information retrieval: A study of user queries on the web," SIGIR Forum 32(1), pp. 5-17, 1998.

[7] C. Silverstein, M. Henzinger, M. Hannes, “Analysis of a very large Web search engine query log”, SIGIR Forum 33(3), pp. 6-22, 2002.

[8] K. Church, B. Smyth, K. Bradley, P. Cotter, "A Large Scale Study of European Mobile Search Behaviour," Intl. Conf. on Human-Computer Interaction with Mobile Devices and Services archive, MobileHCI '08, pp. 13 - 22, 2008.

[9] A. Amandi, D. Godoy, “PersonalSearcher: An Intelligent Agent for Searching Web Pages.”,

SBIA 2000 and IBERMIA 2000. LNCS, vol. 1952, Springer, Heidelberg, pp. 43-52, 2000.
[10] A. Micarelli, F. Gasparetti, F. Sciarrone, S. Gauch, "Personalized search on the world wide web," The Adaptive Web, LNCS 4321, pp. 195-230, 2007.
[11] G. Liu, B. Liu, "Research on Web Search Engine Based on Agent," 1st Intl. Symp. on Pervasive Computing and Applications, pp. 316 - 320, 2006.
[12] Inxight, An Introduction to Inxight: Bridging The Gap Between Search and Actionable Intelligence, *http://www.sap.com, Aug 2008.*
[13] Second Opinion, *http://www.cisco.com/*
[14] Search-Formula-1, 와이즈넛, *http://www.wisenut.co.kr/solutions/*
[15] Nutch, Apache Software Foundation, *http://lucene.apache.org/nutch*
[16] 최창현, 박건우, 이상훈, "지식검색 서비스에서의 소셜 네트워크 기반 영향력 지수 알고리즘," 한국컴퓨터정보학회 논문지 제 14권, 제 10호, 43 - 53쪽, 2009년. 10월.
[17] 문유진, "정보 검색을 위한 숫자의 해석에 관한 구문적·의미적 판별 기법," 한국컴퓨터정보학회 논문지 제 14권, 제 8호, 65 - 71쪽, 2009년. 8월.

저 자 소 개

손 태 식

2000 : 아주대학교 정보 및 컴퓨터공학부 졸업(학사)
2002 : 아주대학교 정보통신전문대학원 정보통신공학과(공학석사)
2005 : 고려대학교 정보보호대학원 정보보호학과(공학박사)
2004 ~ 2005 : Research Scholar, Univ. of Minnesota
2005 ~ 현재 : 삼성전자 Digital Media & Communications 연구소 책임연구원
관심분야 : Wireless/Mobile Network Security, WSN/WPAN, Anomaly Detection/Machine Learning>

박 병 섭

1989 : 충북대학교 컴퓨터공학과 학사
1992 : 서강대학교 전자계산학과 석사
1997 : 서강대학교 전자계산학과 박사
1997 - 2000 : 국방과학연구소 선임연구원
2000 - 2002 ; 우석대학교 컴퓨터교육과 교수
2002 - 현재 : 인하공업전문대학 컴퓨터시스템과 교수
관심분야 : RFID/USN, Zigbee/Bluetooth, Android Platform

최 효 현

1994 : 서강대학교 전자계산학과 학사
1996 : 서강대학교 컴퓨터공학과 석사
2005 : 서강대학교 컴퓨터공학과 박사
2005 ~ 2009 : 삼성전자 통신연구소 책임연구원
2009 ~ 현재 : 인하공업전문대학 컴퓨터정보과 조교수
관심분야 : RFID/USN, 메쉬 네트워크, M2M, 소셜 네트워크