Journal of the Korean Institute of Industrial Engineers
Vol. 39, No. 2, pp. 119-128, April 2013.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2013.39.2.119

<Original Research Paper>

개인화된 웹 검색 순위 생성

강영기 · 배준수†

전북대학교 산업정보시스템공학과

Customized Web Search Rank Provision

Youngki Kang · Joonsoo Bae

Dept. of Industrial and Information Systems Eng. Chonbuk National University

Most internet users utilize internet portal search engines, such as Naver, Daum and Google nowadays. But since the results of internet portal search engines are based on universal criteria (e.g. search frequency by region or country), they do not consider personal interests. Namely, current search engines do not provide exact search results for homonym or polysemy because they try to serve universal users. In order to solve this problem, this research determines keyword importance and weight value for each individual search characteristics by collecting and analyzing customized keyword at external database. The customized keyword weight values are integrated with search engine results (e.g. PageRank), and the search ranks are rearranged. Using 50 web pages of Goolge search results for experiment and 6 web pages for customized keyword collection, the new customized search results are proved to be 90% match. Our personalization approach is not the way that users enter preference directly, but the way that system automatically collects and analyzes personal information and then reflects them for customized search results.

Keywords: Web Search Rank, Customization, PageRank

1. 서 론

사회가 디지털화 되고 정보시스템이 발달함에 따라, 웹의 중요성이 날로 증가하고 있다. 웹(Web)은 무수히 많은 정보들을 기록하고 있으며, 대부분의 사용자들이 구글(Google)이나 네이버(Naver)같은 포털사이트를 이용하여 정보를 검색하고 있다. 포털사이트를 활용하는 사용자는 자신이 원하는 결과들을 리스트 최상위에 보여주기를 희망한다. 그러나 인터넷에서 제공하는 포털 형식의 검색 정보제공 사이트 대부분은 사용자 개개인이 원하는 맞춤형 정보가 제공되지 않는다. 제공되는 정보는 지역별 및 시간별로 분류되고 보편적인 결과의 웹페이지들을 보여줌으로써, 개개인의 사용자가 요구하는 검색결과를 제대로 표현해주지 못하는 단점이 있다.

예를 들어 웹 포털(Google)에서 '스웨이드'를 검색하면, '가수 스웨이드'와 '옷감 스웨이드'에 대한 결과가 혼재되어 있다. 보편적으로 '가수 스웨이드'가 상위랭크에 노출되어 있으며, 표출되는 사이트의 수도 상대적으로 많다는 것을 확인할 수 있다. 웹 포털에서 이러한 결과를 보여주는 이유는 각 정보제공 포털사업자는 단어의 모호함, 중의성 때문에 전 세계적으로 가장 많이 찾는 결과물을 보여줘야 하기 때문이다. 즉, 대부분의 사용자가 '가수 스웨이드'로 표현된 사이트를 자주 클릭하는 경향을 보이기 때문에 상대적으로 많은 웹페이지를 보여주고 있다.

그러나 개인 사용자는 자신만의 원하는 결과를 찾기를 원할 것이다. 만약, "가수 스웨이드"가 아닌, "옷감 스웨이드"를 가장 먼저 찾기를 원할 수도 있다. 하지만 디지털로 되어있는 컴퓨터는 개인적인 고객의 성향을 반영하지 못한다. 이는 자연어의 모호함(동음이의어의 문제) 때문일 수도 있으며, 각각 고객의 성향을 파악해서 DB화하여 제공하기에는 많은 한계를 가지고 있기 때문일 것이다(Brin and Page, 1998). 또한 개인적 성향을 분석하는 저장소를 둔다면 서버 부하가 커질 것이며,

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업지원을 받아 수행된 것임(No. 2010-0025650).
† 연락저자 : 배준수 교수, 561-756 전북 전주시 덕진구 백제대로 567 전북대학교 산업정보시스템공학과, Tel : 063-270-2332, Fax : 063-270-2333, E-mail : jsbae@jbnu.ac.kr
2013년 1월 31일 접수; 2013년 2월 27일 수정본 접수; 2013년 3월 6일 게재 확정.

검색 속도도 느려질 수 있다.

국내에서의 시멘틱 검색의 도입에는 한계가 있다. 시멘틱 검색은 국내 포털 사이트에서도 제공하고 있지만, 그것은 단순히 사용자의 성향을 사용자가 입력해주어야만 그 결과에 대한 관련 정보를 보여줄 수 있다(Lee, 2003). 2012년 현재 전 세계에서 가장 유명한 웹서비스 제공자인 구글(Google.com)도 페이지랭크(Pagerank) 방식을 사용한다(Jung, 2005). 페이지랭크는 웹페이지에서 단어와 외부와 연결된 하이퍼링크를 수집하여 중요한 웹페이지를 선정하여, 원하는 단어를 검색하였을 때, 가장 상위에 위치한 페이지가 표시된다. 페이지 랭크 방식은 매우 좋은 방식이지만 사용자의 개인적인 검색성향을 반영해주지 않는다.

만약, 사용자의 독립적인 데이터베이스를 구축한다면, 개별 사용자의 취미나 특기, 자주 찾는 웹페이지를 분석하여 사용자의 성향을 파악한 후, 사용자가 원하는 단어를 포털에 입력하였을 때, 보편적 결과가 아닌 개인사용자의 취향이 맞는 결과를 보여줄 수 있다면 좋은 검색 서비스가 될 수 있다. 즉 사용자의 성향을 어떻게 입력, 수치화시켜, 그 정보를 바탕으로 어떻게 활용하는가에 대한 방법이 매우 중요하다. 개인이 클릭하여 수집된 키워드 데이터를 계산하여 사용자의 개인성향에 대한 가중치를 정하고, 페이지랭크 결과에 개인성향의 가중치를 더한다면 사용자의 성향을 반영한 검색 순위가 나올 것이다.

본 논문에서는 동음이의어를 포함한 여러 가지 키워드 검색의 중요도의 선정방법을 제안하고 있다. 사용자의 개인적인 성향을 반영하는 데이터베이스를 독립적으로 구축하고, 인터넷 포털에서 제공하는 웹 검색결과에 개인사용자 정보를 추가적으로 반영하여 웹 리스트의 각각의 성향을 재분석하여 포털에 나타난 결과를 재순위화 시키는 방법을 제안한다. 또한, 사용자의 성향을 입력, 수치화 시키며, 그 정보를 바탕으로 어떻게 활용하는가에 대한 방법이 매우 중요할 수 있다. 이는 외부 저장소, 키워드수집, 그리고 페이지랭크 기법을 혼용하는 방식을 사용한다면 해결할 수 있다.

<Figure 1>은 개인 성향을 반영하는 웹 검색의 전체적인 구조도를 나타낸다. 사용자 인터페이스는 구글 검색 결과를 왼쪽에 보여주고, 개인화된 웹 검색 결과를 오른쪽에 보여준다. 또한 구글의 검색과 개인화된 검색의 비중을 가중치로 입력하는 부분이 중간에 있다.

본 논문은 제 2장에서 관련연구를 소개하고, 제 3장은 개인별 키워드의 중요도를 자동 설정하는 방법을 설명하고, 이것을 반영한 웹페이지의 검색 순위 결정은 제 4장에서 정의된다. 제 5장은 제안된 방법론의 실험과 평가가 있고, 마지막으로 결론을 맺는다.

2. 관련 연구

2.1 관련 연구

웹 검색 순위의 효율성을 나타내는 연구는 <Table 1>과 같이 기존에도 활발히 연구되어 왔다. 특히 검색기록 분석을 이용하여 성향정보를 파악하고 이 결과를 다시 검색결과에 반영하여 개인의 관심정보를 효율적으로 제공하는 방안들이 연구

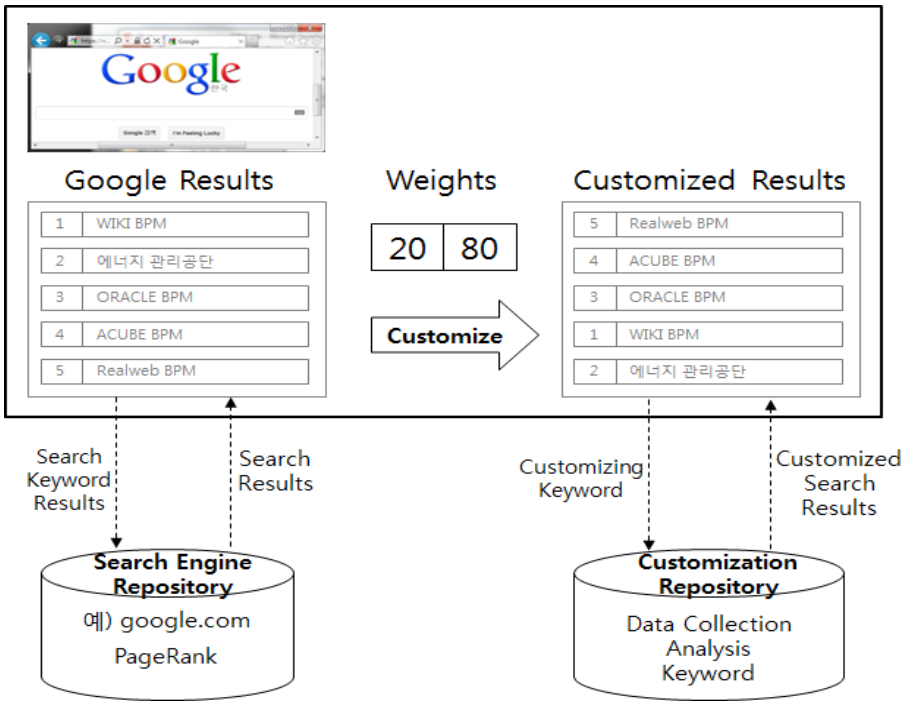

Figure 1. Architecture of customized web search

Table 1. Literature survey of web search

저자	Han *et al*.(2010)	Kim *et al*.(2009)	Jun *et al*.(2002)	본 논문
문제정의	동음이의어에 대해 검색결과를 제대로 반영하지 못하는 문제		정보검색시스템과 질의응답시스템의 문제	동음이의어 문제
분석목적	개인성향을 반영한 동음이의어 판단이 목적		선호 웹 문서의 선택기준 파악	개인 성향을 반영한 동음이의어 해결
기본정보	KISTI글로벌 동향정보리스트	과학기술 학회에서 제공중인 논문 DB	샘플 Q&A정보 리스트	웹 포털의 검색 결과 리스트
분류방법	과학기술표준 코드로 개인성향 분류	서지 DB에 기술된 DDC 코드로 개인성향 분류	직업코드와 관심분야 코드를 분류	검색기록 정보
연구결과	사용자 선호정보 결과 반영 리스트 생성		직업코드에 따른 정확한 Q&A 결과를 표현	웹 포털의 랭킹 결과 재 순위화

되어 왔다(Han *et al*., 2010; Jun *et al*., 2002; Lee and Cheon, 2010). 또한 동음이의어에 대해 검색결과를 제대로 반영하지 못하는 문제들에 대해서 개인코드를 활용하여 분류하는 예시도 제시되었다(Kim *et al*., 2009; Han *et al*., 2010). 정보검색 시스템과 질의응답 시스템의 문제를 선호 웹 문서의 선택기준을 파악하는데 활용하여, 직업과 관심분야 코드를 분류하는 연구도 이루어져 왔다(Kim and Ahn, 2003; Lee, 2010; Jun *et al*., 2002).

하지만 이들 논문에 사용된 기본 데이터는 인터넷 웹 포털의 정보를 사용하지 않고 독자적인 데이터베이스를 사용하였다. 이는 인터넷 웹 포털에 직접 적용이 어려운 문제점을 가지고 있다. 이전의 연구들은 사용자의 개인성향을 반영하는데 있어서, 분류 코드로 코드화시킨 부분이 장점이지만, 본 논문은 웹 포털의 검색결과에 사용자의 키워드의 가중치를 혼합시켜 웹 포털의 검색 순위을 최신화한 면에서 차이가 있다. 종합하면, 본 논문에 제시된 방법이 실제 웹 포털에 적용할 수 있는 실용적인 측면에서 더욱 현실적인 방법을 제시하고 있다

2.2 구글의 페이지랭크 (PageRank)

페이지랭크는 월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법이다. 페이지랭크는 사용자가 링크를 따라 이동하는 것을 시뮬레이션 함으로써 사용자가 각 웹 페이지를 방문할 확률을 구하여 웹페이지의 중요도로 사용하고 있다. 페이지랭크 알고리즘은 수집된 전체 웹 페이지에 대하여 행렬 곱셈을 수렴할 때까지 반복하기 때문에 웹의 크기가 커질수록 많은 시간을 소요한다(Park *et al*., 2011).

2.3 검색 성능측정(정확도와 재현율)

문서의 수는 기하급수적으로 증가해온 반면 사용자가 문서를 찾는 능력은 그렇지 못하다. 사람들은 여전히 검색 결과 중 수개 및 수십 개 정도만을 보려고 한다. 그러므로 문서모음의 크기가 증가함에 따라 높은 정확도를 측정하는 성능척도가 필요해진다. 이중 정확도(Precision)과 재현율(Recall)은 정보검색(Information Retrieval)에서 중요한 성능 측정 기준으로 사용하는 지표이다(Do *et al*., 2003; Yoon, 2009). 사용자가 검색을 할 때, 어떤 특정한 키워드를 준다면 웹에는 그 키워드에 관련된 문서(Relevant document, *R*)와 관련 없는 문서(Irrelevant document, *N*)가 있을 것이다. 만약 "BPM"이라는 키워드로 검색을 한다면, 검색엔진이 찾은 문서들의 집합(Collected document, *C*) 안에는 관련이 있는 문서와 관련 없는 문서가 혼재되어 있을 것이다. 그러면 정확도와 재현율은 다음과 같이 정의된다. $Precision = \frac{|R \cap C|}{|C|}$, $Recall = \frac{|R \cap C|}{|R|}$ 본 논문은 위에서 제시된 정확도와 재현율을 활용하여 제안한 검색 시스템의 성능을 측정해보도록 한다.

3. 개인화된 웹 페이지 검색 순위 생성

본 논문이 제안하는 내용의 전체 구조도는 <Figure 2>와 같다. 사용자가 웹 포털(구글)에 키워드를 입력하여 나온 랭킹페이지의 결과를 외부저장소(Search Engine Repository)에 저장한다. 저장된 값은 하이퍼링크 주소, 랭크, 웹 페이지 키워드들의 웹페이지 정보들로 구성되어 있다. 그리고 사용자가 이전에 인터넷을 통하여 열어보았던 참조 페이지들을 사용자 저장소(Customization Repository)에 저장한다. 참조 페이지를 사용자 저장소에 저장하는 이유는 검색엔진 구글에 표현된 결과값에 사용자의 결과값을 반영하기 위함이고, 저장되는 페이지정보는 구글에서 검색했던 결과와 같은 하이퍼링크 주소, 랭크, 웹 페이지 키워드로 구성되어 있다. 위 2개 저장소의 결과값을 재계산하여 새로운 랭킹을 생성한다. 기존 검색엔진이 보여주는 방법과 다른 점은 사용자의 관심도를 수치화하여 검색엔진에 표현된 웹 리스트를 재구조화 시키는데 장점이 있다. 또한 사용자가 원하는 페이지를 연속적으로 수집함으로써 사용자의 관심도를 동적으로 재구조화시킬 수 있는 장점이 있다.

세부적인 단계는 크게 개인화된 키워드를 생성하는 1단계

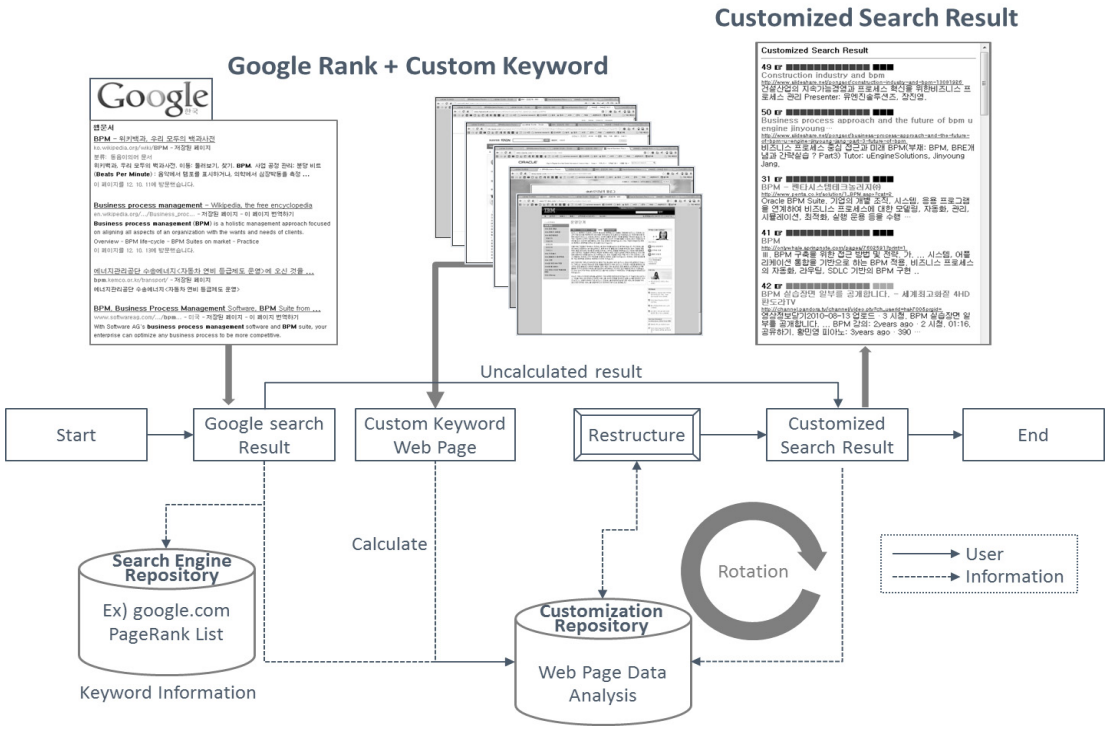

Figure 2. Flowchart of customized web search

와 웹 페이지를 재정렬하는 2단계로 나눌 수 있다. 1단계는 검색엔진의 결과인 페이지랭크(*GR*)를 수집, 페이지랭크 키워드 빈도(*PK*) 측정, 사용자별 선호 키워드 빈도(*CK*), 사용자 선호가 반영된 페이지랭크(*KF*)로 이루어져 있고, 2단계는 개인화 가중치(*CW*)를 결정하는 것과 최종적으로 개인화된 웹 검색 순위(*CR*)를 구하는 것으로 이루어져 있다. 이제부터 한 단계씩 설명한다.

3.1 검색엔진의 페이지랭크 수집(Google Rank, *GR*)

<Figure 3>과 같이 3개의 웹 페이지가 있다고 가정해보자. 각 페이지 이름은 RA, RB, RC이다. 여기서 검색 엔진에서 제공하는 페이지의 순위(*GR*)는{RA = 1, RB = 2, RC = 3}이다.

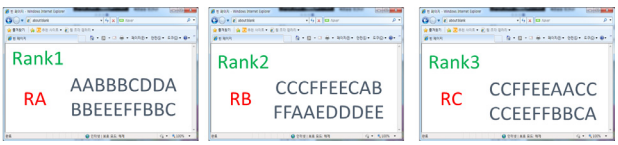

Figure 3. Web page examples

3.2 페이지랭크 결과의 키워드 빈도(PageRank Keyword Frequency, *PK*)

검색 엔진에서 키워드를 입력하였을 때 표시되는 검색 결과의 웹 페이지 리스트를 수집한 후 수집된 웹 페이지들의 키워드들의 빈도를 *PK*라 정의한다. <Figure 3>의 예제 웹페이지를 기준으로 각 웹 페이지의 키워드 빈도(*PK*)를 계산하면 <Table 2>와 같다.

3.3 사용자 검색기록의 웹 페이지 키워드 빈도(Custom Keyword, *CK*)

사용자 참조 웹 페이지를 선정한 후, *PK*를 계산하는 방법과 같은 방법으로 참조 웹 페이지들의 키워드들을 추출하여 빈도가 높은 키워드들을 선정, 사용자 키워드 리스트(Custom Keyword)를 생성한다. 하지만 *PK*는 각각의 웹 페이지의 키워드 빈도를 추출한 반면, *CK*는 참조 문서의 모든 키워드들의 총합에서 가장 빈도가 높은 키워드들을 찾는 부분에서는 차이가 있다. 예를 들어, 사용자가 인터넷을 검색할 때 나타나는 웹 페이

Table 2. Keyword frequency of example web pages

페이지	GR	Text	Keyword Frequency(PK)
RA	1	AABBBCDDABBEEEFFBBC	3A, 7B, 2C, 2D, 3E, 2F
RB	2	CCCFFEECABFFAAEDDDEE	3A, 1B, 4C, 3D, 4E, 4F
RC	3	CCFFEEAACCCCEEFFBBCA	3A, 2B, 7C, 0D, 4E, 4F

지들의 키워드의 빈도(CK)가 [C, E, F]가 높게 나타났다고 가정해보자. 즉 Custom Keyword(*CK*) = {C, E, F}. 그렇다면 [*PK* ∩ *CK*]는 <Table 3>과 같이 표현될 수 있다. 각 페이지별{PK1, PK2, PK3}로 *CK*(Custom Keyword)와 매치된 즉, 키워드{C, E, F}들과 교집합 관계에 있는 키워드들의 수를 각 웹페이지별로 추출한 후, 추출된 키워드들의 전체 개수(빈도)를 모아 PK1, PK2, PK3 등 3개의 웹페이지를 나눠 평균 키워드 수치 11.3을 찾는다.

Table 3. Custom keyword and AVG(*PK*1, 2, 3 ∩ *CK*)

Page Keyword	Frequency	Sum Frequency
*PK*1(RA)∩*CK*	2C, 3E, 2F	7
*PK*2(RB)∩*CK*	4C, 4E, 4F	12
*PK*3(RC)∩*CK*	7C, 4E, 4F	15
AVG(*PK*1, 2, 3∩*CK*)	PK1+PK2+PK3/3pages	11.3

3.4 각 웹페이지의 가중치(Total Keyword Frequency, *KF*)

이전에 구해진 *CK*와 *PK* 값을 이용하여 값을 구한다. *KF* 값은 각 웹페이지의 최종 가중치를 구하는 공식으로 검색엔진에서 랭크된 모든 웹 페이지의 키워드 평균치를 각 웹 페이지에 나타난 빈도값(*CK*∩*PK*)을 나누어 각각의 웹 페이지 가중치를 구한다.

$$KF = \frac{CK \cap PK}{AVG(CK \cap PK)}$$

단, *KF* = Total Keyword Frequency, *CK* = Custom Keyword Frequency, *PK* = Google Keyword Frequency

예를 들어, 각 웹 페이지{RA, RB, RC}의 키워드 빈도에 <Table 3>에서 찾아진 평균값(11.3)을 나누어 *KF*(Total Keyword Frequency) 값을 생성한다. 생성된 *KF* 값은 <Table 4>와 같이 {RA = 0.61, RB = 1.06, RC = 1.32} 값이 생성되었다.

Table 4. Keyword Frequency(KF)

Page Keyword	*PK*/Average	Total Keyword Frequency(*KF*)
RA = 7	7/11.3	0.61
RB = 12	12/11.3	1.06
RC = 15	15/11.3	1.32

3.5 개인화 가중치(Customization Weight, *CW*)

위와 같은 정의는 동음이의어의 문제를 해결 할 수 있는 방법으로 제시되었지만, 어느 정도 비율로 개인화된 결과를 얻을지에 대해 조정할 수 있다. 즉 사용자의 개인성향을 반영하거나 반영하지 않는 방법을 가변적으로 적용하기 위해 *CW*(Customization Weight)를 정의한다. 만약 *CW* 값이 0이면 단순히 검색엔진 구글의 검색결과(*GR*)를 그대로 생성시킬 것이며, *CW* 값이 0이 아니면 *KF* 가중치를 추가적으로 반영시킬 수 있다. 만약 *CW* 값을 1로 설정할 때는 개인화를 100% 반영한다고 볼 수 있다.

3.6 개인화된 웹 검색순위(Customized Rank, *CR*)

$$CR = \frac{1}{GR} + KF \times CW$$

단, *CR = Customized Rank, GR = Google PageRank, KF = Total Keyword Frequency, CW = Customization Weight*

제 3.4절에서 각 웹페이지별로 키워드 중요도 값인 *KF*(Total Keyword Frequency) 값을 생성하였다. 생성된 *KF* 값과 검색엔진 구글에서 페이지랭크를 이용하여 랭킹된 웹 리스트를 함께 반영하기 위하여 *GR*(Google Rank)의 가중치를 더한다. *GR* 값의 생성방법은 검색엔진 구글 페이지에서 키워드를 사용하여 검색하였을 때 최상단에 랭크되어있는 웹페이지를 "1" 그다음을 "2" … 순위로 가중치를 더한다. 위의 방법을 사용하면 랭킹이 낮은 웹페이지가 가중치가 높아지기 때문에 1/*GR* 값을 사용한다. <Table 5>는 *KF* 와 *GR* 값을 같이 반영한 *CR*(Customized Rank) 값의 결과이다. 검색엔진에서 검색된 페이지{RA, RB, RC}의 총 *CR* 값을 계산하여{RC, RA, RB} 순으로 재순위화 순서를 확인할 수 있다.

Table 5. Customized Ranking(CR)

페이지	*KF*	*KF*+(1/*GR*))	*GR*	*CR*
RA	0.61	0.61+1/1 = 1.61	1	**2**
RB	1.06	1.06+1/2 = 1.56	2	**3**
RC	1.32	1.32+1/3 = 1.62	3	**1**

4. 실험결과 비교 분석

실험은 검색엔진 구글에서 "BPM"이라는 키워드를 검색한 결과를 기본 데이터로 선정하였다. 또한 사용자가 참조할 수 있는 웹페이지 6개를 참조페이지로 선정하였다. 구글 검색엔진의 검색결과와 비교하기 위하여 사용자 참조 페이지는 네이버, 구글 등 다양하게 추출되었다. 또한 키워드를 "BPM"으로 선정한 이유는 "BPM"은 다양한 동음이의어를 지니고 있기 때문이다. 데이터베이스는 MS-SQL 2008 Server을 사용하였으며, 언어는 ASP, HTML, Visual Basic 등을 사용하였다.

키워드 "BPM"의 의미는 인터넷에서 다양한 동음이의어를

가지고 있다. 검색엔진 구글에서 “BPM” 키워드로 검색을 실시한 후 랭크된 50개의 페이지를 수집하여 그 의미를 분석한 결과 아래와 같이 4가지의 동음이의어로 분류되었다. 1) Business Process Management 2) Beats Per Minute(Tempo) 3) bpm 156 (Music Band) 4) Hotel BPM(Brooklyn New York에 있는 고급 호텔)

4.1 사용자 참조 페이지 수집

본 실험에서는 <Table 6>과 같이 6개의 사용자 참조페이지를 수집하였다. 위 페이지는 네이버 및 구글에서 수집되었으며, Business Process Management과 관련성이 높은 문서들로 선정하였다. 위 수집된 페이지들은 차후 50개의 페이지와의 빈도분석을 통하여 관련 키워드를 수집할 수 있다.

4.2 검색엔진에 나타난 웹페이지 수집

검색엔진 구글에서 “BPM”으로 검색된 50개의 웹 페이지를 수집하였다. 50개의 웹 페이지에서 수집된 데이터는 *GR*(Google Ranking) 값과 “Business Process Management”와 관련여부, 그리고, 웹 페이지 주소, 웹 페이지의 키워드이다. 수집된 웹 페이지 정보는 <Table 7>에 나타난다. 관련여부는 “Business Process Management”와 관련 있으면 “Y” 관련되어 있지 않으면 “N”으로 정의하였다.

4.3 검색엔진에 나타난 웹 페이지 빈도 측정

키워드 “BPM”으로 입력된 50개의 웹 페이지들은 각 페이지

Table 6. 6 reference web pages selected by user

No	Engine	Page Name	Page Abstract
1	Naver	네이버 지식백과	업무프로세스를 표준화・간소화하고, 비정형화된 업무구성을 시스템화해 …
2		네이버 블로그	비유한다면, BPM(Business Process Management) 제 3의 물결은 목적지까지 이동하는 그 자체라고 할 …
3		네이버 지식인	BPM의 정의와 사례에 대해 화끈하게 알려주세요 ^^ 기업 … BPM(Business Process Management)은 환경에 유연하게 대응하기 위한 프로세스를 관리하기 위한 방안이다.
4	Google	IBM-BPM	오늘날 기업은 새로운 비즈니스 모델을 통해서 시장에서의 경쟁력을 확보할 필요가 있습니다.
5		Oracle BPM	Oracle Business Process Management. 왜 오라클을 선택해야 하는가? 업계 선도적인 BPM. 오라클은 비즈니스 분석가를 위한 모델링 툴, 시스템 통합을 위한 개발자 …
6		BPM-핸디소프트	질문 : 핸디소프트는 BPM 표준과 관련하여 어떤 활동을 하고 있나요? 답변 : 핸디소프트는 다음과 같은 표준화 활동을 하고 있습니다. 1997년부터 WfMC의 Funding …

Table 7. 50 “BPM” related pages searched by Google

GR	Relationship	Page Title
1	n	BPM-위키백과, 우리 모두의 백과사전
2	n	에너지관리공단 수송 에너지<자동차 연비 등급제도 운영>에 오신 것을 …
3	y	Oracle Business Process Management
4	y	BPM-한국IT서비스 산업협회
5	y	ACUBE BPM-SWbiz 소비즈-삼성SDS 솔루션 홈페이지
6	y	BPM Suite-리얼웹
7	y	Social BPM (alpha)-OpenSource BPMS uengine.org
8	y	BPM-IBM
9	n	제닉스의 사고뭉치 : MP3 파일의 BPM 측정 프로그램
10	n	달리고 달리고 또 달려라! BPM 빠르기에 따라 골라 듣는 …
…		…
49	y	Construction industry and bpm
50	y	BPM-IBM

Table 8. *PK*(PageRank Frequency) of 50 web pages

F : Frequency

Page1	F	Page2	F	Page3	F	Page4	F	Page50	F
문서	3	공인연비	6	Oracle	19	프로세스	9	PaaS	21
문서는	3	자동차	4	BPM	16	지원	8	Social	16
이	3	2010	2	위한	6	기능	5	Business	12
10월	2	검색	2	제품	6	기반	4	Open	12
BPM	2	원	2	Data	3	BPM	3	Platform	12
다른	2	자동차	1	Process	3	다양한	3	개발	12
동음이의어	2	1,825.38	1	Suite	3	반도체	3	시장	11
모든	2	10/07～10/13	1	문의	3			클라우드	11
보기	2	111.3	1	비즈니스	3	업무	3	BPM	10
수	2	114.82	1	솔루션	3	있는	3	Process	9
있습니다.	2	2,009.72	1			제공	3	공개	9

만의 고유한 키워드를 가지고 있다. 각 웹페이지들의 키워드의 빈도를 분석한 결과인 <Table 8>은 각 페이지의 랭킹과 단어의 빈도(F)를 요약하였다.

4.4 참조할 웹페이지의 빈도분석

Table 9. *CK*(Custom Keyword Frequency) of 6 reference web pages

No	6 pages	Frequency
1	BPM	32
2	프로세스	22
3	Oracle	19
4	SOA	17
5	비즈니스	17
6	Process	11
7	[BPM]	9
8	업무	9
9	IBM	8
10	프로세스의	8
…		

"BPM"의 키워드 가중치를 정의할 6개의 참조 웹 페이지를 앞서 설명하였다. 참조할 웹 페이지는 모든 참조 웹 페이지들의 키워드를 수집하여야 한다. 때문에 <Table 9>는 6개의 참조 웹페이지의 모든 키워드의 빈도를 분석한 결과이다.

4.5 각 웹 페이지의 총빈도값 분석(Total Keyword Frequency, *KF*)

각 웹페이지의 *KF*(Total Keyword Frequency)값을 구한다. 본 실험에서는 AVG($CK \cap PK$)의 값이 5.060으로 나타났다. 위 평균값을 이용하여 각 웹페이지의 *KF* 값을 계산한 값은 <Table 10>에 나타나있다.

$$KF = \frac{CK \cap PK}{AVG(CK \cap PK)}$$

4.6 각 페이지의 CR(Customized Ranking)값을 분석하여 랭킹을 재순위화

이번 단계는 *CR*(Customized Ranking) 값을 구한다. <Table 11>은 기존의 *GR* 값과 재순위화 된 랭크 *CR* 값이다. 1/*GR*을

Table 10. *KF*(Keyword Frequency)

No	CK∩PK	KF	Web Page Title
1	4	0.791	BPM-위키백과, 우리 모두의 백과사전
2	0	0.000	에너지관리공단 수송 에너지<자동차 연비 등급제도 운영>에 오신 것을 …
3	10	1.976	Oracle Business Process Management
4	0	0.000	BPM-한국IT서비스 산업협회
5	10	1.976	ACUBE BPM-SWbiz 소비즈-삼성SDS 솔루션 홈페이지
…			
49	16	3.162	Construction industry and bpm
50	15	2.964	Business process approach and the future of bpm u engine jinyoung …

Table 11. *CR*(Customized Web Rank)

GR	$CK \cap PK$	*KF*	$\frac{1}{GR}$	$\frac{1}{GR}+(KF \times CW)$	*CR* (Ranking)
1	4	0.791	1.000	1.791	14
2	0	0.000	0.500	0.500	32
3	10	1.976	0.333	2.310	8
4	0	0.000	0.250	0.250	40
5	10	1.976	0.200	2.176	9
6	5	0.988	0.167	1.155	18
7	5	0.988	0.143	1.131	19
8	1	0.198	0.125	0.323	38
9	4	0.791	0.111	0.902	21
10	8	1.581	0.100	1.681	15
				…	
49	16	3.162	0.020	3.182	1
50	15	2.964	0.020	2.984	2

사용한 것은 검색엔진 구글에서 제공하는 가중치를 그대로 활용하기 위함이다.

5. 구현 및 평가

본 논문에서 제시하는 기법이 기존검색결과보다 검색에 얼마나 더욱 효율적으로 표현되는지를 측정하기 위해 상위에 제시된 결과들을 이용하여 새로운 랭킹리스트를 생성해 보았다. 이 결과는 단순하고 간단하게 구축될 수 있으며, 이전의 결과값보다 사용자의 개인 성향을 더욱 반영한 결과값을 상위에 노출시키는 결과를 가져왔다.

<Figure 4>의 왼쪽에 있는 결과는 검색엔진 구글의 검색결과이며, 오른쪽에 있는 것은 재설계된(CR) 검색결과이다. 검색결과에 나타나 있는 흰색 바와 회색 바는 사용자가 찾고자 하는 “BPM”(Business Process Management)과 웹페이지의 내용이 일치하는가를 나타낸 것이다. 흰색 바는 다른 동음이의어를

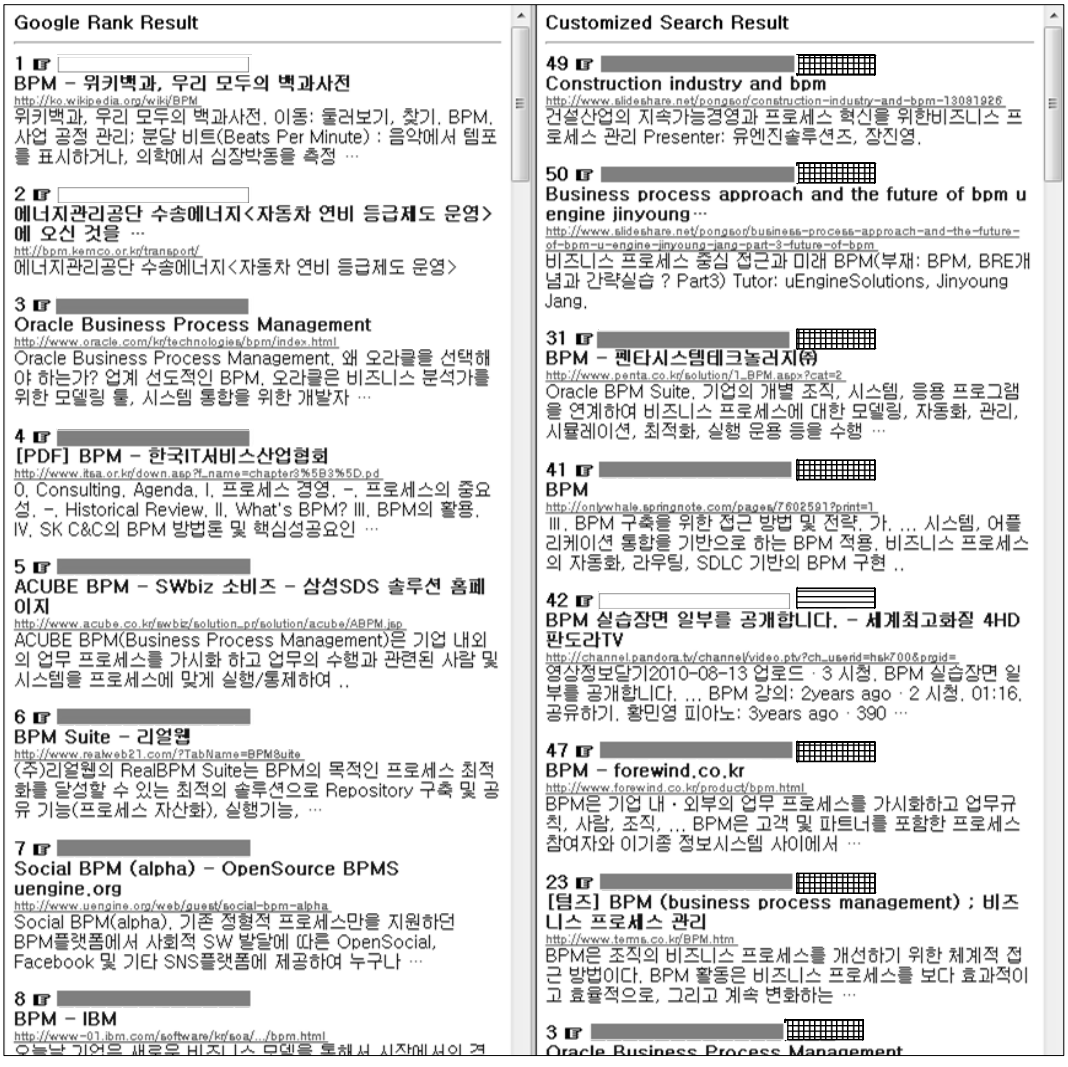

Figure 4. Customized web search result implemented by ASP

표시한 것이고, 회색 바는 “BPM”(Business Process Management)와 의미론적으로 일치하는 페이지를 보여준다. 또한 <Figure 4>의 오른쪽 *CR*(Customized Ranking)에 있는 격자무늬와 가로무늬의 의미는 구글에 있는 BPM 속성 페이지와 계산식으로 판단하였을 때 판단한 BPM 속성페이지가 맞는지 틀린지의 여부를 AVG(*CK*∩*PK*)값으로 판단한 것이다. 예를 들어 AVG가 5.060이라고 계산되었을 때 “BPM”(Business Process Management) 빈도가 5.060 이상일 때 BPM 페이지가 회색 바로 표시되어 있으면 격자무늬, 또한 빈도가 5.060 이하일 때 BPM 페이지가 흰색 바로 표시되어 있으면 격자무늬를 나타낸다. 반면에 AVG(*CK*∩*PK*)가 빈도값이 5.060 이상인데 “BPM”(Business Process Management)페이지가 흰색이면 가로무늬로 표시되며, 빈도값이 5.060 이하인데 페이지가 회색이면 가로무늬로 표시된다. 이와 같은 결과를 통하여 각 페이지들의 키워드 평균값을 기준으로 웹페이지의 성향이 원하는 성향인지 자동으로 판단하는 방법을 제시하고 있다.

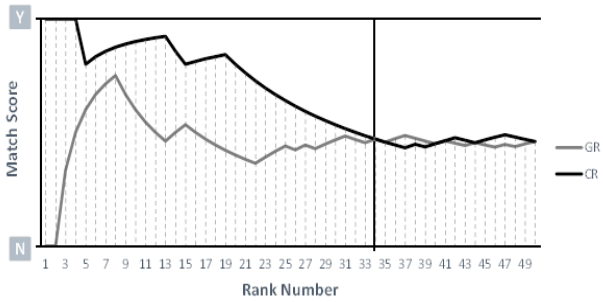

Figure 5. Cumulative graph of match score

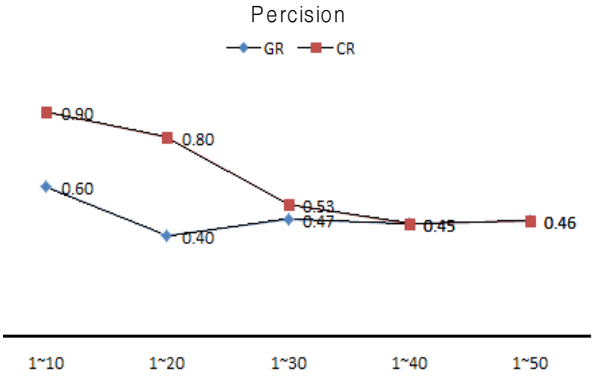

Figure 6. Precision evaluation

<Figure 4>의 결과를 <Figure 5> 누적결과 그래프로 나타낼 수 있고, 구체적으로 성능평가 지수를 도입하여 정확도와 재현율을 <Figure 6>과 <Figure 7>로 나타낼 수 있다. 각 그림은 랭크 10단위의 누적값과 분할값을 가지고 있다. 기준을 상위 20개로 가정하였을 때 정확도는 기존보다 200% 향상됨을 확인할 수 있었으며, 상위 10개의 재현율은 150% 향상된 것을 확인할 수 있었다.

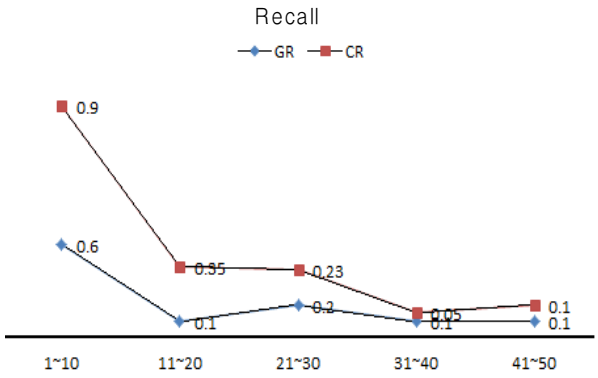

Figure 7. Recall evaluation

<Table 12>는 개인 성향 가중치(*CW* 값)의 변화에 따라 관련성 있는 문서들이 어떻게 랭크되는지를 보여준다. 0으로 표기된 값은 사용자와 관련성 없는 문서이며, 1로 표기된 값은 사용자와 관련성 있는 문서이다. *CW* = 0일 때 결과는 이전에 *GR* 값과 동일한 것으로 나타나지만, *CW* = 0.1로 10%만 반영되어도. 사용자의 성향에 맞게 변화됨을 확인할 수 있다. *CW* = 1 값은 사용자 가중치를 100% 반영한 결과이다. *CW* = 0.5 이상부터 관계없는 웹 문서의 랭크가 고정되거나 줄어들면서 랭킹값이 점점 나아지고 있음을 확인할 수 있다.

마지막으로 사용자가 원하는 웹 페이지를 제대로 판별하였는지를 체크하였다. 사용자가 원하는 웹 페이지 인지 여부를 판별하는 방법은 사용자 키워드 빈도의 판별값 공식이다. 본 실험에서 “BPM”으로 검색한 사용자 키워드 집합의 평균값은 5.020이다. 만약, 각 랭킹에 표현된 웹 페이지의 키워드가 사용자 평균 값인 5.020을 넘으면 “A”이며, 5.020을 넘지 않으면 “B”로 판별한다. 그리고 이전에 각 웹 페이지의 속성을 판별한 결과인 “0”과 “1” 값을 비교하였을 때 서로 “1” = “A”이거나 “0” = “B”일 때 격자무늬로 표시하고 “1” = “B”이거나 “0” =

Table 12. Precision and Recall according to *CW* change

Top 10’ Precision/Top 30’ Recall											
CW	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Recall	0.47	0.60	0.57	0.57	0.53	0.53	0.53	0.53	0.53	0.53	0.53
Recall(%)	0%	29%	21%	21%	14%	14%	14%	14%	14%	14%	14%
Precision	0.60	0.70	0.80	0.80	0.80	0.80	0.80	0.80	0.90	0.90	0.90
Precision(%)	0%	17%	33%	33%	33%	33%	33%	33%	50%	50%	50%

"A" 처럼 자동 계산 값과 판별 값이 다를 때는 가로무늬로 표시하였다. 이와 같은 작업을 반복 해본 결과 총 50개의 문서 중 40개(약80%)가 격자무늬로 표기되어 웹페이지에 사용자의 성향을 추론할 때 빈도의 평균 값을 사용하는 것이 효과적이라는 것을 확인할 수 있었다.

6. 결론 및 향후 연구 방향

현재 포털 서비스는 일반적인 키워드로 검색하였을 시에는 정확한 결과값을 보여주는 반면, 동음이의어 같은 중복된 의미의 키워드에 대해서는 적절한 결과값을 보여주지 못한다. 때문에 동음이의어의 키워드를 입력하면 모든 검색사용자가 선택하는 보편적인 결과의 웹 리스트만 보여줄 뿐 개인사용자 각각의 맞춤 검색결과를 제공하지 못하는 단점이 있다. 특히 동음이의어를 완벽하게 해결하지 못하는 단점이 있다. 본 논문은 이를 해결하기 위해 페이지랭크 기법을 활용하는 검색엔진인 구글의 검색결과에 개인 성향 정보를 추가하여 검색엔진에서 보인 웹 랭크를 재구조화시켜 사용자가 원하는 검색결과를 재반영시키는 개인화된 웹 검색결과를 생성하였다.

본 논문에 제시한 방법은 사용자의 성향을 반영하는 검색결과를 효과적으로 생성하는 장점이 있다. 그리고 동음이의어의 키워드로 검색한 페이지를 사용자가 원하는 페이지로 재정렬시켜 사용자에 맞는 웹 페이지 검색 결과를 향상시켰다. 또한 웹 포털 서비스를 대상으로 하기 때문에 높은 활용도가 예상되며, 사용자의 개인성향 가중치를 조정하여 검색엔진의 사용방법을 사용자가 원하는 값으로 조정할 수 있는 장점이 있다.

추후에는 사용자 데이터베이스에 개인화 키워드를 추가하여 개인성향의 검색 정확도 더욱 향상시킬 수 있는 방법을 찾아내야 하며, 다양한 개인성향 키워드 생성방법을 개발 및 적용하며, 검색결과에 대한 개별적인 사용자의 만족도를 어떠한 방식으로 분석해야할 것인지는 추후에 연구해야할 과제이다.

참고문헌

Brin, S. and Page, L. (1998), The anatomy of a large-scale hypertextual Web search engine, *Journal of Computer Networks and ISDN Systems*, **30**(1～7), 107-117.

Do, H. H., Melnik, S., and Rahm, E. (2003), Comparison of schema matching evaluations, *Web, Web-Services, and Database Systems, Lecture Notes in Computer Science*, **2593**, 221-237.

Han, H.-J., Kim, J.-S., Lee, S.-H., Choe, H.-S., Kim, K.-Y., and You, B.-J. (2010), Search Result Personalization using Search History Analysis, *Journal of Korean Society for Internet Information*, **10**, 125-126.

Jun, B.-H., Kim, J.-H., and Kwak, H.-Y. (2002), Design and Implementation for User Oriented Search System using the Information of History, *Journal of research institute of advanced technology*, **10**(1), 91-97.

Jung, B.-J. (2005), A study on satisfaction index of internet portal site : with emphasis on internet user behaviors, environments and demographic characteristics, Department of Management Korea National Open University.

Kim, K.-Y., Shim, K.-S., and Kwak, S.-J. (2009), A Personalized Retrieval System Based on Classification and User Query, *Journal of the Korean Library and Information Science Society*, **43**(3), 163-180.

Kim, H.-H. and Ahn, T.-K. (2003), An Experimental Study on the Internet Web Retrieval Using Ontologies, *Korea Society for Information Management*, **20**(1), 417-455.

Lee, J. H. (2003), Ontology Languages for the Semantic Web, *Korea Information Science Society review*, **21**(3), 18-27.

Lee, J.-H. and Cheon, S. H. (2010), Re-ranking for Search result using association relationship and TF×IDF, *Korean Institute of Information Scientists and Engineers*, **37**(1), 349-352.

Lee, S.-J. (2010), Analysis of Preference Criteria for Personalized Web Search, *The Journal of Korean association of computer education*, **13**(1), 45-52.

Park, S.-J., Lee, S.-H., and Hwang, D.-H. (2011), A Web Contents Ranking System using Related Tag and Similar User Weight, *Journal of Korea Multimedia Society*, **14**(4) 567-576.

Yoon, S. H. (2009), Using Query Word Senses and User Feedback to Improve Precision of Search Engine, *Journal of Information Management*, **26**(4), 81-91.