가중치 기반 웰빙식품 정보 검색 시스템☆

Weight-based Wellbeing Food Retrieval System

편 광 범* 윤 은 일** 류 근 호***
Gwang Bum Pyun Unil Yun Keun Ho Ryu

요 약

건강에 대한 관심이 높아지면서 웰빙 관련 정보의 필요성이 중요해졌다. 웰빙 정보검색은 인터넷 검색 엔진이나 블로그, 개인 홈페이지 또는 대중매체를 이용한다. 하지만, 웰빙 식품에 관한 정보는 구하기 어렵다. 그래서 검색엔진은 웰빙식품에 대한 정보검색이 필요하게 되었다. 본 논문은 가중치기반의 웰빙식품 검색엔진을 설계하고 구현한다. 수많은 페이지를 탐색해 웰빙 식품 키워드가 포함되어있으면 이것을 식별하여 가중치를 추가하는 방식이다. 사용자가 키워드를 이용하여 검색하면 웰빙 관련 페이지가 우선적으로 나올 수 있게 구현했다. 웰빙관련 식품의 식별에 사용되는 키워드들은 사전형식으로 되어있다. 그래서 삽입, 삭제, 수정이 가능하다. 역 파일은 직접파일인 해싱 방식으로 저장한다. 본 논문의 엔진을 이용하여 성능평가를 한 결과 웰빙식품 키워드에 대하여 타 검색엔진에 비해5~15%의 향상된 결과를 보였다. 본 논문에서는 검색엔진의 설계방식과 웰빙식품에 특화된 랭킹선정방식을 제안한다.

ABSTRACT

As the interests in health grow higher, necessity of Well-being relation informations get more importance. We get the information of well-being, tinternet retrieval system or blog, homepage and media. Although, it is not easy to find informations of well-being food. So, retrieval system has been requiring information about well-being food. In this paper, Weight-based Wellbeing Food Retrieval System is designed and implemention. Finding numerous pages and if well-being keywords includes page, it was identified and add weight. User searching for keywords, it implement, well-being food pages comes at the first. Keywords for discrimination makes type of dictionary, so it can insert, delete, modify. Inverted files saves hasing(direct-based file). Retrieval System in this paper is experimental result, at keywords of well-being food show 5~15% imprement than another Retrieval System. In this paper, Weight-based Wellbeing Food Retrieval System's designed and proposed way to raking for well-being food.

☞ KeyWords : Wellbing, Food, Wellbing-Food, Search Engine, Additional Weight Ratio, Rangking,
웰빙, 식품, 웰빙식품, 검색 엔진, 가중치, 랭킹

1. 서 론

최근 웰빙이 확산되면서 소비자의 건강식품 및 안전성에 대한 관심이 높아지고 이와함께 웰빙식품과 관련된 정보 제공의 필요성이 강조되고 있다. '2008 한국의 사회지표'[1]의 부분별 사회지표 표6-11에 따르면 일반인들의 건강에 대한 관심은 2006년 51.7%에서 2008년 57.5%로 증가했다.

TV와 인터넷 페이지 등의 각종 매체를 통해 다

* 준 회 원 : 충북대학교 컴퓨터과학과 석사과정
pyungb@chungbuk.ac.kr
** 정 회 원 : 충북대학교 전자정보대학 컴퓨터전공 조교수
yunei@chungbuk.ac.kr(교신저자)
*** 정 회 원 : 충북대학교 전자정보대학 컴퓨터전공 교수
khryu@dblab.chungbuk.ac.kr
[2009/12/08 투고 - 2009/12/15 심사(2010/03/08 2차) - 2010/05/12 심사완료]
☆ 이 논문은 2010년도 정부 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. : 2010-0004197)" 그리고 "이 논문은 2010학년도 충북대학교 학술연구지원사업에 의하여 연구되었음(This work was supported by the research grant of the Chungbuk National University in 2010)". 또한, "이 논문은 2010년 교육과학기술부로부터 지원받아 수행된 연구임" (지역거점연구단육성사업 / 충북BIT연구중심대학육성사업단)

양하고 많은 건강 정보가 제공되고 있지만 이것은 상황에 알맞는 정보가 아닌 브로드 캐스팅 방식의 정보전달이다. 원하는 정보를 얻기 위해서는 인터넷을 이용해야 하는데 이 경우에 네이버 카페나(cafe.naver.com) 다음 카페(cafe.daum.net)등 포털 사이트에서 제공하는 카페를 이용하거나 웰빙 관련 상업용 사이트에서 얻는 것이 대부분이다. 이곳에서는 원하는 정보의 일부만 얻을 수 있고 인터넷에 퍼져있는 수많은 자료들을 한번에 확인하는것도 쉽지 않다. 또한 우리 먹거리에 대한 불신과 정보의 부재가 이슈가 되면서 웰빙관련 식품에 대한 검색사이트의 필요성이 더욱 부각되고 있다[2]. 현재 웰빙 식품관련에 대한 정보를 검색할 수 있는 사이트로 조사에 따르면[3] 소비자의 인지도를 가진 사이트는 공공기관 9.5점 연구기관 8.5점 학술협회 7.9점 병의원 7.8점 순이고 개인 페이지와 검색 사이트는 평균 6.3점으로 점수가 매우 낮은 수준이었다. 식품관련 정보를 쉽고 정확하게 찾을 수 있는 방안으로 웰빙 관련 단어에 대한 키워드로 검색된 웹페이지 중에서 웰빙관련 자료가 많은 페이지를 찾아주는 정보검색 형태가 있으며 또 다른 형태로는 커뮤니티 형식으로 블로그나 홈페이지를 제작하여 사용자가 직접 정보를 작성하는 방식이 있다. 전자의 경우 인터넷상에 퍼져있는 수많은 자료들을 대상으로 하지만 잘못된 자료가 검색되거나 필요하지 않은 정보를 걸러 내는 것이 쉽지 않다. 후자의 경우 정확한 자료와 자료 제작자와의 커뮤니티 연결이 매우 용이하지만 자료의 양이 한정 되어있고 많은 사용자의 참여가 필요하다. 본 논문에서는 전자의 경우 즉, 인터넷 검색으로 수많은 페이지를 가져와 랭킹을 매겨 사용자가 원하는 자료를 쉽게 구할 수 있도록 설계 했다. 본 논문에 기술된 검색엔진의 특징으로 웰빙 관련 페이지인 경우 추가로 가중치를 부여하여 관련 페이지가 좀 더 높은 랭킹 점수를 얻는 방식으로 되어있다. 추가 가중치에 필요한 웰빙 관련 키워드는 사전 형식으로 관리자가 직접 추가, 삭제, 변경을 할 수 있고, 검색되는 페이지는 웰빙 관련 자료가 가장 위쪽으로 올 수 있도록 구현했다. 본 논문은 2장에서 웰빙에 대하여 조사하고 어떠한 부분이 중요한지, 그리고 기존검색 시스템에 대해서 살펴보고 3장은 시스템 랭킹선정 방식 4장은 실제로 시스템이 어떤 방식으로 구성되어있는지 5장은 시스템의 성능을 평가했다. 마지막으로 6장은 결론과 향후 연구이다.

2. 관련 연구

2.1 웰빙이란?

웰빙이란 오늘날 '행복, 삶의 만족, 질병이 없는 상태를 모두 포괄하는 개념'으로서 현대 산업사회의 병폐를 인식하고, 육체적・정신적 건강의 조화를 통해 행복하고 아름다운 삶을 영위하려는 사람들이 늘어나면서 나타난 새로운 삶의 문화 또는 그러한 양식을 말한다. 1980년대 중반 유럽에서 시작된 슬로푸드(slow food) 운동, 1990년대 초 느리게 살자는 기치를 내걸고 등장한 슬로비족(slow but better working people), 부르주아의 물질적 실리와 보헤미안의 정신적 풍요를 동시에 추구하는 보보스(bobos) 등도 웰빙의 한 형태이다. 이와같은 웰빙에 관한 내용 에서 본 논문은 웰빙식품에 초점을 맞추었다. '식생활관련 웰빙지향 소비가치와 웰빙식품 소비행동'[4] 에 의하면 가정식을 즐기는 소비자가 많아 식재료의 중요성이 요구되며, 식재료 중에서도 생선과 유기농산물에 많은 소비형태가 나타나므로 이를 고려하여 검색엔진을 구성한다.

2.2 기존 웰빙관련 검색 엔진

현재 웰빙 관련 검색 엔진으로 개발된 엔진은 크게 부족하다. 하지만 이것을 대체할 수 있는 주제별 검색 엔진으로서 네이버와 구글등 대부분의 검색엔진이 채택하고 있다. 이와같은 상업용 포털 사이트는 영화 음악 인물 쇼핑등 여러 가지 많은

종류의 주제별 검색을 지원하고 있지만 웰빙과 식품에 관련된 주제별 검색이 없다. 일반 상업 사이트로 웰빙 식품관련 검색을 하려면 일반검색에 의존해야한다. 종종 특정키워드에서는 식품관련보다 타 분야의 페이지가 더 랭킹이 높아 원치 않는 정보를 얻는 경우가 많았다. 필요한 웰빙 식품 관련 페이지를 찾기 위해 많은 페이지들을 찾아야 했다. 공공기관에서 자료를 찾는 방법도 있지만 이 자료들은 사전적 의미로서의 자료들이 대부분이기에 신문기사나 블로그 같은 더 많은 정보를 얻을 수 있는 기회가 없다. 요즘 사회적으로 관심이 높아지고 있는 웰빙식품 관련 중요도를 고려하여 보다 정확한 웰빙식품 검색 및 분석이 필요하다[5]. 이를 이용하여 검색 시스템을 제작하면 웰빙 식품 정보에 대해 사용자의 기대치가 증가한다.

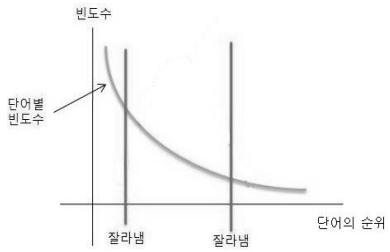
3. 랭킹선정 방식

3.1 문서내 페이지 필터링

우선 문서내 주제어가 되는 키워드를 찾아 분류 한다. 본 논문에서는 확률적 통계 기법이 사용되었다. 이와 관련된 연구로는 Naive Bayes 방식이 있고 Zipf's law의 방법이 있다.[6, 7] 본 논문에서는 Zipf's law를 이용하여 빈도수가 너무 많거나 적으면 필터링 할 수 있도록 만들었다. Zipf's law란 많이 사용되는 단어들이 존재하며 매우 드물게 사용되는 단어들을 모아보면 많은 수를 차지하고 있다. 빈도수를 F, 순위를 X라 할 때 아래와 같이 표현된다.

$$F(X) = c^* X^{(-a)}$$

c와 a는 상수가 되며 a는 1에 근접한 값을 가진다. 빈도수와 순위의 곱은 상수c에 항상 일정하게 되어 빈도수가 높을수록 낮은 순위를 갖게 된다.



(그림 1) Zipf's law를 활용한 필터링

여기에 너무 높은 순위와 너무 낮은 순위를 잘라내는(cutoff)방식을 취한다.[7] 자동색인의 통계적기법과 한국어문헌의 실험[8]에 의하면 한국어 문헌의 색인어 선정실험에서 30%에 가까운 비율로 비 주제어가 된다. 표 1의 알고리즘에 의해 돌출된 점수는 0~1000의 값을 가지게 되는데 비 주제어의 비율인 30%를 잘라내어(cutoff) 검색엔진의 효율을 향상시킨다. 이렇게 cut off된 빈도수와 역문헌빈도 가중치와 결합하게 된다.

(표 1) zipf's law를 활용한 pruning 알고리즘

활용 알고리즘 : zipf's law

```
//키워드 리스트중에서 가장 큰 빈도수를 갖는 값을 찾는다
average_frequency = find_average(word_list);
for(i=0;i<all_word_num;i++)
{
    //빈도수가 평균 빈도수보다 크거나 같으면 높을수록 점수가 감소한다
    if(word_list[i].frequency >= average_frequency)
    {
        // 랭킹점수를 구한다 비율로 점수를 구한다
        word_list[i].zip_score = 1000 -
        (word_list[i].frequency /
        (average_frequency*2)*1000);
    }
    //빈도수가 평균 빈도수의 작으면 높을수록 점수가 증가한다
    if(word_list[i].frequency < average_frequency)
    {
        // 랭킹점수를 구한다 비율로 점수를 구한다
        word_list[i].zip_score =
(word_list[i].frequency
        / (average_frequency*2)*1000);
    }
     // 300점 이하로 받은 키워드를 cutoff 시킨다
    if(word_list[i].zip_score < 300)
            word_list[i].score = 0;
}
```

3.2 웰빙 페이지 선별

페이지 내부에 웰빙 식품관련 키워드가 나오면 이 페이지는 웰빙관련 페이지로 식별되어 추가점수인 W를계산하게 된다. 계산 방식은 I를 페이지 내부의 단어들이라 하고 If 는 I단어의 빈도수 weight는 I가 포함된 카테고리 사전의 가중치다.

$$W = \frac{\sum_{I=0}^{n} I_f \times weight(I)}{n}$$

예를 들어 페이지에 사과, 우유, 고등어라는 키워드를 가지고 있고 사과의 빈도수는 10 우유는 7 고등어는 3 이고 카테고리별 가중치는 표 2와 같다. 사과의 weight는 순위3의 채소, 과일류가되어 33%인 0.33이 되고 우유의 weight는 순위2의 유제품, 음료, 과자와빵이 되어 36%인 0.36이 되고 고등어의 weight는 순위1의 잡곡, 생선류 39%인 0.39가 되어 계산되는 웰빙식품 관련점수 W = (10 * 0.33) + (7 * 0.36) + (3 * 0.39) / 3 = 2.33이 된다.

(표 2) 카테고리별 추가 점수

순위	카테고리	추가 비율
1	잡곡, 생선류	39%
2	유제품, 음료, 과자와빵	36%
3	채소, 과일류	33%
4	육류	25%
5	나머지 식재료	20%
-	일반 키워드	0%

(표 3) 웰빙 페이지 식별 알고리즘

```
int make_page_add_score
(page *word_list, int all_word_num)
{
    int score=0;
     for(i=0;i<all_word_num;i++) {
          // 첫 번째 가중치 사전 로드
           if(cereals_and_fish_dic(word_list.[i].word)) {
                score = FIRST_CATEGORY
                break;
           }
          // 두 번째 가중치 사전
           if(milk_and_drink_dic(word_list.[i].word)) {
                score = SECOND_CATEGORY
                break;
           }
          // 세 번째 가중치 사전
           if(vegetable_and_fruit_dic(word_list.[i].word)) {
                score = THIRD_CATEGORY
                break;
           }
          // 네 번째 가중치 사전
           if(meat_dic(word_list.[i].word)) {
                score = FOURTH_CATEGORY
                break;
           }
          // 다섯 번째 가중치 사전
           if(etc_dic(word_list.[i].word)) {
                score = FIFTH_CATEGORY
                break;
           }
     }
     return score / all_word_num;
}
```

사용자들의 성향에는 일반적인 패턴이 존재한다. 일반적인 웰빙식품에 관한 검색만을 제공하는 것이 아닌 소비자의 웰빙식품 성향 패턴을 조사하여 높은 성능과 사용자의 만족을 얻을수 있다[9]. 만 20세 이상의 기혼여성 소비자를 대상으로 웰빙 지향 식재료를 조사한 결과[4, 10] '잡곡, 생선류', '유제품, 음료, 과자와 빵', '채소, 과일류', '육류' 순으로 소비형태가 나타났다. 추가점수에 들어가는 수치는 조사서의 응답 평균치를 기준으로 부여하였다. 이 평균치에 의하면 복수 선택 설문지에서 각 항목을 체크한 비율을 나타낸 것이다. 잡곡, 생선류를 선택한 사람들은 39%이고 유제품, 음료 과자와 빵은 36%, 채소, 과일류는 33%, 육류는 25%를 선택하였다. 그 외로 다른 종류의 식품을 선택한 사람은 20%이다. 설문지에 나온 수치와 같은 가중치를 주면 웰빙식품 페이지를 찾을 수 있다.

이런 방식으로 소비자들이 많은 관심을 가지고 있는 내용부터 순서대로 점수를 높게 줄 수 있다. 이런 방식으로 랭킹점수를 만들면 웰빙식품 관련 페이지가 좀 더 많은 점수를 받아 쉽게 찾아볼 수 있다. 이와 같은 카테고리별 로 사전 형태로 구성하여 언제든지 추가 삭제 변경이 가능하도록 구현했다. 또한 웰빙식품 선호도에 관한 소비자의

관심도가 바뀌어도 간단한 수작업을 통해 변경 가능하도록 구성 했다. 기본이 되는 웰빙식품사전은 foodnara.go.kr에서 제공하는 사전을 사용했다.

3.3 최종 랭킹점수

단어 빈도수가 F가되고 이것과 전체 페이지의 개수 N, 단어 k를 포함하고 있는 페이지의 수 dk를 이용하여 역 문헌빈도 가중치를 계산하면 아래와 같다[11, 12, 20].

N : 문헌집단에 포함된 문헌의 총 수
d_k : 용어 k를 포함하고 있는 문헌의 수

$$idf = \log_2(N/d_k) + 1$$
$$= \log_2 N - \log_2 d_k + 1$$

전체 페이지 N을 용어 k를 포함하는 문헌의 수를 나누어 해당 키워드의 가중치를 계산한다.[12] 여기에 페이지에서 가지고 있는 웰빙관련 단어의 점수 W를 추가 하면 최종적으로 나오는 랭킹 R은 다음과 같이 확정된다.

$$R = F \times idf \times W$$

여기서 나온 최종 랭킹점수 R의 내림차순으로 웹 서버에서 페이지 리스트가 출력된다. W는 찾은 웰빙식품 페이지 추가 가중치다. W를 통해 다른 페이지들과 차별을 두었다.

(표 4) 최종 랭킹점수 알고리즘

활용 알고리즘 : idf

```
//웰빙관련 추가 키워드 가중치
page_add_score =   make_page_add_score
                    (word_list, all_word_num);
for(i=0;i<all_word_num;i++){
      //idf 적용 idf = log2N-log2Dk+1
      word_list[i].idf_score = log2(all_page_num) -
      log2(count_word_page_num(word_list[i].word))+1;
      //최종 랭킹점수 적용 R = F * idf * W
      word_list[i].score = word_list[i].zip_score
      * word_list[i].idf_score
      * page_add_score;
}
```

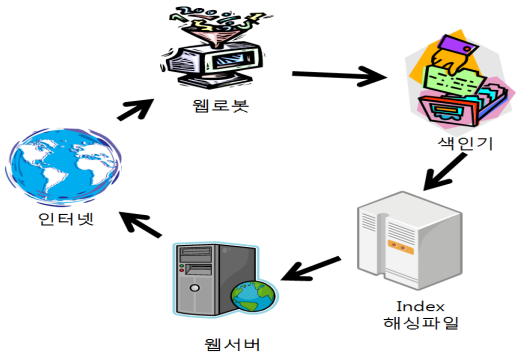
4. 엔진의 구조

본 시스템은 웹페이지 수집기인 웹로봇(Crawler)와 형태소 분석과 색인, 역파일을 만들어 주는 인덱서(Indexer) 그리고 역파일을 검색하여 페이지로 보여주는 웹서버(WebServer)로 구성되어있다[13]. 언어는 웹로봇(Cralwer)과 인덱서는 C로 제작되었고 웹서버는 PHP로 제작되었다. 효율적인 시스템 설계를 위해 클래스간 결합 척도를 참고해 재사용성을 고려했다[14].

4.1 웹로봇(Cralwer)

웹로봇(Cralwer)은 웹사이트의 링크들을 돌아다니면서 웹페이지를 수집하는 역할을 한다. 웹로봇은 여러 형태의 웹로봇이 있으며 웹 문서 변경 예측을 하여 페이지 재수집 기간을 장비와 상황에 맞추어 하는 것이 좋다[15]. 웹로봇으로 부터 수집된 웹페이지는 파일형태 또는 이진 데이터 형태로 저장된다. 이 검색엔진에서의 웹로봇은 씨드페이지(Seed page)[1]를 가장먼저 방문후 그곳에 링크된 웹페이지를 차례로 방문하여 가져온다[16, 17]. 그 방법은 가져온 웹페이지를 파싱하여 링크(html 태그의 href[2]) 부분을 가져온다) 부분을 인식하여 가져온 링크들을 큐에 저장한다. 큐에 저장된 링크 중 접속이 불가능한 링크(자바스크립트를 이용한 링크나 플래시를 이용한 링크인 경우)를 선별하여 제거한다. 그 후 방문한 페이지를 수집 날자와 함께 페이지 주소와 함께 이진데이터 형태로 인덱서(Indexer)로 전송한다. 이 웹로봇(Cralwer)은 C로 제작되었으며 다른 웹로봇보다 다른 점은 html태그 제거 모듈이 추가되었다. 이를 이용하여 텍스트 정보만이 인덱서(Indexer)로 전송한다.

1) 웹로봇에서 가장 처음으로 방문하는 페이지 또는 사이트
2) HTML 태그에서 다른 주소나 페이지로 링크하는 명령어



(그림 2) 시스템 구조

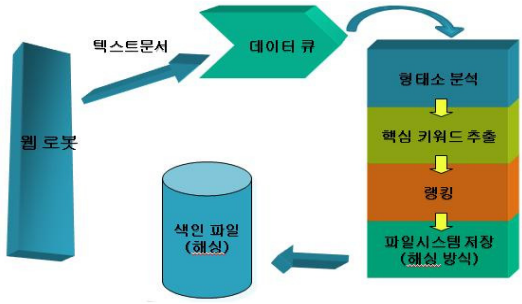
4.2 형태소 분석기

형태소 분석기는 텍스트 문서에서 키워드를 추출한다. 웹로봇(Cralwer)에서 전송된 텍스트 데이터를 단어 사전파일에 등록된 단어를 기준으로 단어를 추출하여 빈도수와 함께 저장한다. 본 형태소분석기는 국민대학교 언어공학 정보검색 연구실에서 개발한 것으로 기본적으로 영어와, 한국어 형태소를 분석하여 출력한다. 특수 기능으로는 정보검색 시스템을 위한 자동색인 기능, 한글 맞춤법 검사 및 교정 기능, 한국어 복합명사 분해 기능, 한글 문장의 자동 띄어쓰기 기능, 사전기능을 이용한 불용어와 시소러스[3] 기능이 있다. 본 검색 시스템에 사용된 기능은 영어와 한글형태소 분석기능, 불용어와 시소러스 기능이다. 본 형태소 분석기의 분석 방식은 전처리 모듈에서 부호나 숫자 같은 필요 없는 부분을 제거하고 분석을 용이하게 하기 위해 조합형 코드로 변환 시킨 후 토큰별로 단어를 나눈다. 분석모듈은 토큰별로 나눠진 각 단어를 사전을 이용하여 체언분석, 용언분석, 단일형태소 분석을 거치게 된다. 합성명사인 경우에는 복합명사 추정을 통해 구별하며 줄임말이나 미등록 단어가 있는 경우 사전에 추가하는 방식으로 되어있다. 마지막으로 후처리 모듈에서 단어들을 정리하여 출력하게 된다. 빠르고 효율적으로 형태소 분석을 한다[18].

3) 어휘(語彙)를 뜻의 관점에서 분류하여 체계화한 것.

4.3 인덱서(Indexer)

형태소 분석기에서 추출된 단어(keyword)와 빈도수를 가지고 이것을 주어진 공식에 따라 랭킹 점수를 부여 한다. 랭킹 구현부분은 페이지 내부 필터링, 전체 페이지에 대한 역 문헌도 가중치 랭킹, 웰빙식품에 대한 가중치 추가 부분으로 나누어져 있다. 페이지 정보와 키워드별 랭킹점수를 역파일로 만든다. 역파일은 해시 방식으로 파일시스템에 저장되게 되며 단어(keyword)와 빈도수, 단어의 위치(location), 웰빙페이지 랭킹 점수를 가진다.



(그림 3) 인덱서(Indexer)

4.4 역파일

검색된 정보를 저장하는 역파일은 해시 방식으로 파일시스템으로 표 5와 같은 방식으로 저장한다. 가장 앞에 단어를 구분하는 키워드가 오며 다음에 빈도수, URL[4], 웰빙식품관련 랭킹점수, 미리보기를 위한 긁어온 텍스트파일을 명시하는 해시 코드로 이루어져 있으며 각각 애트리뷰트[5]마다 TAB으로 구분한다. 역파일에 적용되는 해시저장 방식은 키워드의 아스키 코드값을 수치형 변수 형태로 변환하여 해시값으로 사용한다. '우산'이라는 단어 아스키 코드 값이 '3DEF 5A2E' 라 하면 해시값은 '3DEF5A2E' 라는 값으로 만들어

4) 웹 문서의 각종 서비스를 제공하는 서버들에 있는 파일의 위치를 표시한다.

5) 데이터테이블에서 종류를 구분하는 항목

지게 되는데 문자변수의 경우는 1바이트 크기의 데이터만 저장할 수 있기 때문에 1바이트씩 가장 오른쪽바이트부터 분리해 자리수 별로 더해주는 방식으로 구현하면 위에 설명한것과 동일한 해시값을 구현할 수 있다.

Hashcode : 해시값
ascii code : 각각 문자마다 아스키 코드값
location : 키워드에서 특정 문자의 위치
Hashcode = Hashcode + (ascii code * location)

(표 5) 역파일 내부구조

Key word	Frequency	URL
Ranking Score	Text-file Hash-code	END

이렇게 만들어진 해시값으로 파일을 하나씩 만들어 역파일을 생성하게 된다. 이런 방식으로 파일을 생성하면 키워드 마다 하나의 역파일이 생성되기 때문에 많은 역파일이 생기게 된다. 그래서 운영체제의 파일시스템 관리상 성능저하가 발생한다. 그에 대한 해결책으로 하나의 디렉터리(폴더)에 파일이 집중되는 현상을 피하기 위해 1000개의 파일마다 디렉터리를 만들어 구분했다.

4.5 웹서버

웹서버에서는 사용자에게 웹 인터페이스를 제공하며 질의를 처리하여 관련 정보를 보여주는 역할을 한다. 사용자가 원하는 키워드를 입력하면 이 키워드를 간단한 연산을 이용해 여러 키워드로 만들어 낸다. 예를 들면 '삼겹살 상추' 키워드가 입력되면 '삼겹살' 키워드와 '상추' 키워드를 나누어 해당 키워드가 들어있는 역 파일을 접근한다. 두 키워드가 모두 나오는 페이지들은 먼저 보여주고 가장 첫번째 키워드에 대한 검색결과를 다음으로 보여준다. 표 6과 같이 여러 키워드에대한 AND연산의 결과물이 가장 먼저 나오고 각각 순서대로 정렬된다.

(표 6) 웹서버의 쿼리 응답 결과 소팅 방식

QUERY : A, B, C, D
A AND B AND C AND D
A AND B AND C
A AND B AND D
A AND C AND D
B AND C AND D
A AND B, A AND C, A AND D,
B AND C, B AND D, CAND D
A, B, C, D

(표 7) 해시코드생성 & 파일저장

```
for(i=0;i<n;i++){
   //너무 작은 점수를 받은 키워드를 잘라낸다
   if(term[i].score<200||(strlen(ls+tern[i].offset))==1||

(strlen(ts+term[i].offset))==2||(strlen(ts+term[i]>10)
          continue;
   strcpy(word,ts+term[i].offset);
   for(j=0;j<(strlen(word));j++) {
      // hash값을 만든다
      hashnum = hashnum+abs((int)word[j])+(256*j);
      //char -> int 형변환시 절대값을 취해준다
      //(음수일 경우 정확한 해시값을 위해)
      directorynum = hashnum;
      // hash값을 이용해 1000단위로 버킷을 나눈다
   while(directorynum>=1000){
      dir = dir*1000; directorynum = directorynum-1000;
   }
   itoa(dir,dirc,10);
   // 역파일이 저장될 파일의 경로와 이름 생성
   strcat(newpath,"./index/"); mkdir(newpath);
   strcat(newpath,"dirc"); stract(newpath,"/");
   mkdir(newpath); strcat(newpath,path);
   strcat(newpath,".txt");
   //파일출력
   indexfp = fopen(newath,"a+"); feek(indexfp,OI,SEEK_SET);
   fpirntf(indexfp, "%s\t%u\t%s\t%u\t%s\n",
      word,term[i].tf,url,term[i].score,fname);
   strcpy(newpath,""); fclose(indexfp);
}
```

표 6 방식에 따라 키워드에 대한 결과를 순차적으로 보여주는 방식이다. 최종 랭킹 점수가 매겨진 페이지들을 페이지 마다 10개씩 묶어서 보여준다. 랭킹점수가 가장 높은 페이지가 가장 윗부분에 출력되며 내림차순으로 점수가 낮은 순서

대로 정렬 후 출력한다[19]. 또한 단순히 링크 주소만 보여주는 방식이 아닌 역 파에 들어 있는 특정 페이지에 해당 키워드가 어느 위치에 있는지 파악하여 그 주변 200자 텍스트 미리보기 형태로 제공하는 인터페이스를 가지고 있다.

5. 성능분석

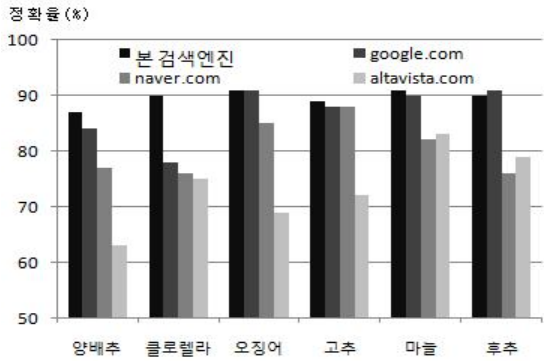
검색 시스템의 정확율을 평가하기위해 다음과 같은 평가 방식을 취하였다. 이 웹문서들은 특정 시드사이트(seed site) 없이 무작위 페이지들에 대해 평가가 이루어졌고 수집된 페이지수는 18만개이다. 평균 정확율(Average-Precision) 방식으로 정확율(Precision)[21]과 재현율(Recall)을 측정했다. 성능평가를 시행한 실험환경은 하드웨어 부분으로 2.8Ghz의 CPU와 2Gbyte RAM, 소프트웨어 부분으로 OS는 MS Windows XP SP3 가 사용되었고 웹서버로는 Apache2가 사용되었다.

(표 8) 실험 환경

H/W	CPU	Pentium IV 2.8Ghz
	RAM	DDR2 2G Bytes
	HDD	250G Bytes
S/W	OS	MS Windows XP SP3
	Web Server	Apache 2
	Language	C, PHP

5.1 타 사이트와의 정확율(Precision) 평가

식품관련 검색 사이트를 찾았지만 이에 특화된 검색사이트가 거의 전무한 관계로 일반 상업용 검색사이트 www.naver.com, www.google.co.kr, kr.altavista.com를 대상으로 비교를 했으며 본 연구의 주제를 얼마나 충실히 이행했는지 알아보기 위해 100개의 검색결과에 대한 정확율을 식품관련 키워드대해 평가했다. 특정 키워드를 검색하였을 때 상위 50개의 페이지를 가지고 이것이 웰빙 식품관련 페이지인가 판단하여 평가 했다. 기존 검색 시스템은 일반적인 주제를 가지고 랭킹을 계산하기 때문에 웰빙 식품관련이 아니더라도 관련이 높은 경우 상위 페이지로 보여주는 경향이 있다. 본 검색 시스템은 웰빙 식품에 초점을 맞추어 개발되었기에 다른 상용 검색 사이트에 비교하여 적은 페이지 수로도 심사용 논문 그림 4와 같이 향상된 결과가 나왔다.



(그림 4) 정확율 평가

5.2 평균정확율(Average-Precision)

평균 정확율은 여러 쿼리에 대한 평균 정확율을 측정하는 평가이다. 특정 개수의 적합한 문헌을 찾을때까지 몇 개의 페이지가 검색되었고 여러 쿼리를 측정하여 그 평균을 구한다. 표 9에의 평균 정확율 표에서 Q는 쿼리, 수준은 적합한 문헌의 개수이며 안의 내용은 수준의 문헌을 찾기까지 검색된 페이지의 개수다. 수준40에서(적합한 문서가 40개가 될 때까지) Q1은 42개의 페이지가 검색되었고 그 비율은 95%이다. 평균은 여러개의 쿼리에 대한 검색된 페이지의 평균을 나타낸다.

(표 9) 평균 정확율

단위 : 개

수준	Q1	Q2	Q3	평균
20	20(100%)	20(100%)	20(100%)	21(100%)
40	42(95%)	40(100%)	44(91%)	42(95%)
60	66(91%)	66(91%)	65(92%)	65(92%)
80	87(92%)	89(90%)	86(93%)	87(92%)
100	112(89%)	116(86%)	108(92%)	112(89%)

웰빙 식품관련 키워드를 이용한 평균 정확율은 표 9와 같다. 평균 정확율에 대한 평가로서 90% 이상의 좋은 결과가 나타났다.

(표 10) joins.com에 게재된 '호박'관련 웰빙식품 뉴스 (2009.12~2010.2)

연번	기사 제목
1	'신이 내린 몸매' 제시카 고메즈, 한국 게임 여전사 된다
2	[용인맛집] 싱싱한 해물의 향연, 해물전문점 '주문진 해물'
3	신 김치 별미 요리
4	[헬스코치] 전립선암 예방 식품, 토마토만 아세요?
5	[백년명가] ① 한 입 베어물면 만두소 '툭'…손맛에 반했다
6	입소문난 동네 떡집의 비결
7	한눈에 멋이, 한입엔 맛이 … 디저트 매력에 푹 빠졌네
8	간편하게 즐기는 전통 디저트
9	[스타쿡 ②] 먹어도 붓지 않는 '호박라면'·한입에 예쁘게 '조랭이떡 떡볶이'
10	건강한 재료로 차린 한 그릇 밥
11	삼청동 카페거리 大탐방 1
12	[food&] 그때는 멀건국·양배추김치 … 지금은 스파게티·오리고기
13	긴긴 겨울밤, 간단간식
14	[골프장 맛&멋] 강추위 녹이는 얼큰한 국물
15	아이를 위해 아빠가 읽어야 할 책
16	[임정식 셰프의 비밀 레시피] 밀라노 해물찜
17	그림책 읽은 후 요리 놀이
18	트렌디 식당에서 찾은 건강 아이디어, 채식 레시피
19	새해소망 다이어트! 30일 식단 플랜
20	[백년명가 ③] 약수로 지어 푸른 솥밥을 먹어보자
21	파워블로거의 힘, 매출 3배 급증한 웰빙 간식
22	'카페 오시정'의 한국식 브런치 레시피
23	[푸드& ①] 새해 소망 생각하는데…뱃속에선 꼬르륵 꼬르륵
24	[임정식 셰프의 비밀 레시피] 시푸드 샐러드
25	연말모임, 세트메뉴로 실속… 경품 이벤트는 보너스
26	+3cm & -3kg, 아이를 위한 식단
27	밥보다 반찬을 많이 먹는다? 식사습관 Q & A
28	구석구석 서래마을
29	전문가 "소녀시대 식단 1,200kcal 섭취? 영양적으로 불균형"
30	[백년명가 ③] 갈비, 참숯으로 구워야 제 맛!
31	'퓨전 떡볶이' 떡볶이의 과감한 변신
32	[food&] 즐겁게, 쉽게, 우아하게 셰프 배틀 3인이 제안하는 4코스 성탄 요리
33	피자, 칼로리가 높아서 고민? 피자랑쥬로 칼로리는 낮추고 영양은 UP!
34	크리스마스 특별한 추억 만들기
35	소녀시대 식단 공개 '하루 고작 800kcal 섭취 대단해'
36	자동차·눈사람·통나무·은종 …색다른 분위기 만들어 볼까
37	소녀시대 식단, 하루에 고작 800kcal? "너무해!
38	이웃사촌과 함께하는 이층집 크리스마스
39	조선시대 122세 할머니의 장수비법
40	[Food&] ②광화문 맛집엔 특허받은 맛에서 보양식까지 다양
41	[Food&] ① 이벤트 가득한 광화문거리, 내 입에도 '맛 이벤트' 터진다
42	자연과 닮은 도시락 문화를 제안하다
43	제시카 고메즈, 김치 다이어트 화제
44	남·여탕 매일 바뀌는 곳, 그 묘미 느껴 보셨나요
45	[헬스코치-木] 혈당·콜레스테롤도 낮춰준다! 고혈압 예방 위한 최고의 식품
46	'공룡나라 쇼핑몰' 매출 4억

5.3 재현율(Recall)

중앙일보 인터넷 신문 사이트 joins.com에서 최근 3개월간(2009년 12월1일~2010년 3월 1일) 웰빙 관련 페이지중에서 식품 '호박' 과 관련있는 뉴스 기사는 46개이다. 표 1에있는 기사들은 내용에 호박과 관련된 내용이지만 제목이 호박과 관련이 없을 수도 있다. 좋은 검색 시스템이라면 46개에 근접한 결과가 나와야 한다. 본 검색 시스템과 한국에서 가장 인기있는 포털 사이트인 naver.com의 검색 재현율 비교를 한 결과 naver.com은 총 46개 중 10개(21.7%)의 관련 뉴스가 검색되었다. 본 검

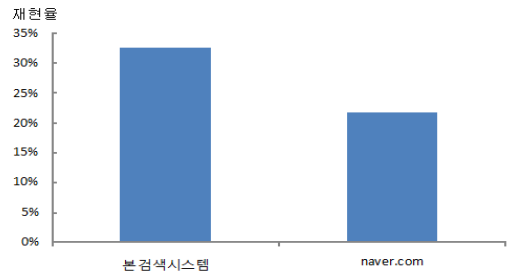
색시스템은 총 46개중 15개(32.6%)의 관련 뉴스가 검색되었다. 표 10을 표본집합으로 본 검색 시스템과 naver.com의 재현율을 비교한 결과 naver.com보다 5개 뉴스를 더 많이 찾아내어 11%가 향상된 모습을 볼 수 있다.

(표 11) naver.com에서 검색된 '호박'관련 웰빙식품 뉴스 (2009.12~2010.2)

연번	기사 제목
1	남·여탕 매일 바뀌는 곳, 그 묘미 느껴 보셨나요
2	[용인맛집] 싱싱한 해물의 향연, 해물전문점 '주문진 해물'
3	입소문난 동네 떡집의 비결
4	[헬스코치] 전립선암 예방 식품, 토마토만 아세요?
5	간편하게 즐기는 전통 디저트
6	[골프장 맛&멋] 강추위 녹이는 얼큰한 국물
7	파워블로거의 힘, 매출 3배 급증한 웰빙 간식
8	[food&] 즐겁게, 쉽게, 우아하게 셰프 배틀 3인이 제안하는 4코스 성탄 요리
9	크리스마스 특별한 추억 만들기
10	[헬스코치-木] 혈당·콜레스테롤도 낮춰준다! 고혈압 예방 위한 최고의 식품
10 / 46(21.7%)	

(표 12) 본 검색시스템에서 검색된 '호박'관련 웰빙식품 뉴스 (2009.12~2010.2)

연번	기사 제목
1	파워블로거의 힘, 매출 3배 급증한 웰빙 간식
2	입소문난 동네 떡집의 비결
3	간편하게 즐기는 전통 디저트
4	[헬스코치-木] 혈당·콜레스테롤도 낮춰준다! 고혈압 예방 위한 최고의 식품
5	[food&] 즐겁게, 쉽게, 우아하게 셰프 배틀 3인이 제안하는 4코스 성탄 요리
6	[임정식 셰프의 비밀 레시피] 밀라노 해물찜
7	+3cm & -3kg, 아이를 위한 식단
8	소녀시대 식단, 하루에 고작 800kcal? "너무해!
9	제시카 고메즈, 김치 다이어트 화제
10	새해소망 다이어트! 30일 식단 플랜
11	'퓨전 떡볶이' 떡볶이의 과감한 변신
12	조선시대 122세 할머니의 장수비법
13	건강한 재료로 차린 한 그릇 밥
14	트렌디 식당에서 찾은 건강 아이디어, 채식 레시피
15	남·여탕 매일 바뀌는 곳, 그 묘미 느껴 보셨나요
15/46(32.6%)	



(그림 5) 재현율 평가

6. 결 론

본 논문은 웰빙 시대에 맞춘 웰빙 식품검색을 정확하고 손쉽게 할 수 있도록 지원해주는 검색 엔진을 설계하고 구현했다. 중요한 점은 웰빙식품 페이지에 대한 가중치를 부여해 사용자가 필요한 정보를 쉽게 검색하도록 랭킹을 구성하게 되어 이를 향상시켰다. 페이지내부의 빈도수와 역 문헌 빈도 가중치를 이용한 랭킹시스템에 웰빙 식품 관련 페이지에 대해 더 높은 가중치를 주게 되어 같은 점수의 페이지라도 웰빙에 관한 내용이 있으면 점수를 가산하여 웰빙식품 관련 페이지를 정확하고 손쉽게 검색할 수 있도록 했다. 성능평가 결과로 웰빙식품 관련 키워드에 대해서 5~15% 향상된 결과를 보여주었다. 추후 연구로서 수집하는 페이지를 대량으로 만든 시스템을 구성하고 성능을 향상시키는 것이다. 검색에 필요한 페이지 수를 늘려 상용 검색사이트보다 뛰어난 검색 시스템을 개발하는 것이 목표다.

참 고 문 헌

[1] 통계청(2008) 2008 한국의 사회지표.

[2] 곽창근, 장종근, 웰빙식품산업 활성화 방안 - 신선편의식품 시장을 중심으로 -, 식품산업과 영양, 제13권, 제1호, pp.17-27, 2008.

[3] 강혜경, 강명희, 유경혜, 이선영, 인터넷 영양정보의 모니터링-메타데이터의 분석, 한국 영양학회지, 제37권, 제8호, 한국 영양학회, 2004년, pp688-700.

[4] 제미경, 전향란, 식생활관련 웰빙지향 소비가치와 웰빙식품 소비행동, 대한가정학회지, 제45권, 제9호, 2007년, pp.63-74.

[5] Thomas Roelleke, A frequency-based and a poisson-based definition of the probab ility of being informative, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp.227-234, July 28-August 01, 2003.

[6] 이재문, 빈발단어집합을 이용한 NaiveBayes의 정확도 개선, 한국인터넷정보학회논문지, v.7, no.3, pp.169-178 , 2006

[7] Carlo Altamirano, Alberto Robledo, "Generalized thermodynamics underlying the laws of Zipf and Benford", in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol.5, pp.2232-2237, 2009

[8] 정영미, 이태영, 자동색인의 통계적기법과 한국어문헌의 실험, 한국문헌정보학회지 제9권, 99~118p, 1982. 12

[9] 구태완, 홍성준, 이광모, 웹 서비스에서 개인 성향 모듈의 설계 및 구현, 한국인터넷정보학회논문지, v.10, no.4, pp.161-176, 2009

[10] 이영민, 백수련, 박홍주, 심근섭, 이희주, 전혜경, 인구사회학적 특성에 따른 웰빙식품관련 정보의 요구도, 한국지역사회생활과학회지, 제17권, 제3호, 2006년 9월, pp.175-182.

[11] Cambridge University Press Information Retrieval, 'Inverse document frequency', 2008.

[12] Akiko Aizawa, An information-theoretic perspective of tf−idf measures, Information Processing and Management: an International Journal, v.39 n.1, p.45-65, January 2003.

[13] 김혜영, 최신성 가중치를 고려한 검색 모형에 대한 연구, 연세대학교 문헌정보학과 대학원, 2007.

[14] 최미숙, 효율적인 시스템 설계를 위한 클래스 간의 결합 척도, 한국인터넷정보학회논문지, v.9, no.5, pp.85-97, 2008

[15] 김성진, 웹 문서 변경 예측, 한국인터넷정보학회논문지, v.8, no.4, pp.149-158, 2007

[16] 윤보현, 구조화된 웹 문서에 대한 자동 정보 추출, 한국인터넷정보학회논문지, v.6, no.3, pp.129-145, 2005

[17] 정창후, 최윤수, 진두석, 김진숙, 윤화묵, 대용량 XML 문서의 효율적인 검색과 관리를 위한 SCOF 모델, 한국인터넷정보학회논문지, v.9, no.1, pp.103-113, 2008

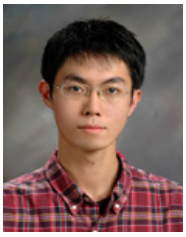
[18] 국민대학교 언어공학 정보검색 연구실, 강승식, KLT2008 형태소 분석기, http://nlp.kookmin.ac.kr/HAM/kor/index.html.

[19] F. Qiu and J. Cho, 'Automatic Identification of User Interest For Personalized Search", In Proceedings of the 15th international conference on World Wide Web, pp. 727-736, 2006.

[20] Donald Metzler, Generalized Inverse Document Frequency, Conference on Information and Knowledge Management, pp. 399-408, 2008

[21] 윤성웅, 채진기 이상훈, 질의 내부 단어 인접도를 이용한 검색 효율 향상 기법, 정보과학회논문지 : 데이터베이스, 제 35권, 제 2호, 192-198p, 2006.

◐ 저 자 소 개 ◑



편 광 범

2010년 충북대학교 컴퓨터공학전공 졸업(학사)
2010년~현재 충북대학교 컴퓨터과학과 석사과정
관심분야 : 데이터마이닝, 정보검색, 데이터베이스
E-mail : pyungb@chungbuk.ac.kr

윤 은 일

1997년 고려대학교 이학석사
1997년~2006년 한국통신 멀티미디어연구소 전임/선임 연구원
2005년 Texas A&M Univ. 공학박사
2005년~2006년 Texas A&M Univ. 포스닥연구원
2006년~2007년 한국전자통신연구원, 선임연구원
2007년~현재 충북대학교 전자정보대학 컴퓨터전공 조교수
관심분야 : 데이터마이닝, 정보검색, 데이터베이스
E-mail : yunei@chungbuk.ac.kr

류 근 호

1976년 숭실대학교 공학사
1980년 연세대학교 공학석사
1980년~1983년 한국전자통신연구원 연구원
1983년~1986년 한국방송통신대학교 조교수
1988년 연세대학교 공학박사
1986년~현재 충북대학교 전자정보대학 컴퓨터전공 교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오인포매틱스
E-mail : khryu@dblab.chungbuk.ac.kr