Journal of Digital Contents Society Vol. 11 No. 3 Sep. 2010(pp. 357-365)

개인화 웹 검색 시스템 기반의 문서 요약 시스템

김동욱*, 강수용**, 김한준***, 이병정****, 장재영*****

요 약

개인화 웹 검색 시스템은 사용자의 검색의도에 따라 질의어 확장, 검색 결과의 재순위화 등의 방법을 통하여 사용자에게 개인화된 검색 결과를 제공한다. 이를 위해 검색 시스템은 질의어와 사용자의 프로파일 정보를 활용하여 사용자의 검색 의도를 파악하고 분석하여, 검색 결과 페이지에 반영하여 보여주게 된다. 이때 검색 결과 페이지는 문서의 URL과 문서의 제목, 작은 텍스트 조각을 표시한다. 여기서 작은 텍스트 조각은 검색 질의어가 포함된 문서의 요약이며, 스니펫이라고 알려져 있다. 사용자는 이러한 문서의 요약을 통하여 웹 문서가 자신이 원하는 정보를 가진 문서인지를 판단하거나, 해당 URL에 직접 접속하지 않고도 원하는 정보를 얻을 수 있게 된다. 따라서 문서 요약은 사용자가 문서를 볼 것인지 아닌지에 대한 중요한 판단 기준이 되며, 만약 문서 요약 시스템이 개인화된 요약 결과를 제공한다면 사용자의 만족도는 더욱 증가할 것이다. 본 논문은 전체 웹 검색 시스템에서 검색 속도의 큰 하락 없이 사용자의 만족도를 증가시킬 수 있는 개인화 문서 요약 시스템을 제안한다.

A Document Summary System based on Personalized Web Search Systems

Dongwook Kim*, Sooyong Kang**, Hanjoon Kim***, Byungjeong Lee****, Jae-young Chang*****

Abstract

Personalized web search engine provides personalized results to users by query expansion, re-ranking or other methods representing user's intention. The personalized result page includes URL, page title and small text fragment of each web document. which is known as snippet. The snippet is the summary of the document which includes the keywords issued by either user or search engine itself. Users can verify the relevancy of the whole document using only the snippet, easily. The document summary (snippet) is an important information which makes users determine whether or not to click the link to the whole document. Hence, if a search engine generates personalized document summaries, it can provide a more satisfactory search results to users.

In this paper, we propose a personalized document summary system for personalized web search engines. The proposed system provides increased degree of satisfaction to users with marginal overhead.

Keywords : summary, snippet, personalization, information retrieval

1. 서론

현재의 검색엔진을 이용하는 사용자들은 다수의 웹 문서로부터 원하는 정보를 얻기 위해 해당 정보를 포함하는 웹 문서를 얻을 수 있는 질의어를 선택하여 검색엔진에 전송하게 된다. 전송된 질의어를 통해 검색엔진은 사용자의 검색 의도를 파악하고 분석하여, 검색 의도와 연관이 높은 순서대로 문서의 URL과 문서의 제목, 작은 텍스트 조각(fragment)을 보여주게 된다. 여기서 작은 텍스트 조각은 검색 질의어가 포함된 문서

※ 제일저자(First Author) : 김동욱
접수일:2010년 09월 07일, 수정일:2010년 09월 29일, 완료일:2010년 09월 30일
* 한양대학교 전자컴퓨터통신공학과
eliudkim@hanyang.ac.kr
** 한양대학교 컴퓨터공학부(교신저자)
*** 서울시립대학교 전자전기컴퓨터공학부
**** 서울시립대학교 컴퓨터과학부
***** 한성대학교 컴퓨터과학부
▣ 본 연구는 서울시 산학연협력사업 (과제번호: NT080624, 연구과제명: 집단지성 기반 사용자 인식 웹검색 시스템 개발) 및 2010년도 한성대학교 교내연구비의 지원에 의해 수행되었음

의 요약이며, 스니펫(snippet)이라고 알려져 있다. 사용자는 이러한 문서의 요약을 통하여 웹 문서가 자신이 원하는 정보를 가진 문서인지 아닌지를 판단하거나, 해당 URL에 직접 접속하지 않고도 원하는 정보를 얻을 수 있게 된다. 따라서 스니펫은 해당 URL에 접속하여 전체 문서를 볼 것인지 않을 것인지에 대한 판단 기준이 되기 때문에 검색 결과의 만족도에 큰 영향을 미친다[1]. 예를 들어, 검색 결과 문서에 해당 내용이 포함되지 않은 요약 정보가 포함되거나, 해당 문서와 크게 관계가 없는 요약 문장을 참고하여 검색 의도와 관련이 없는 문서를 사용자가 접속하게 되는 경우 사용자의 검색 만족도에 나쁜 영향을 미치게 된다.

다수의 검색엔진에서 어떤 방법으로 스니펫을 선택하여 보여주는지는 명확하게 알려지지 않았지만, 해당 문서에서 질의어가 포함된 내용 중 중요한 문장 일부를 보여준다. 하지만 이러한 방법은 웹 문서에 질의어가 포함된 문장이 많은 경우 각 사용자의 선호도에 관계없이 모든 사용자에게 동일한 요약 결과를 보여주게 되어 모든 사용자의 검색 만족도를 만족시켜 주지 못하게 된다. 문장 요약 정보에도 각 사용자의 선호도에 따른 개인화된 검색 결과를 보여준다면 더욱 좋은 검색 결과를 얻을 수 있을 것이다. 하지만 각 문서의 결과 페이지는 사용자가 질의어를 전송하고 결과를 보기 위한 대기시간이 일정시간을 초과하게 되면 오히려 검색 품질의 만족도가 올라가더라도 전체적인 검색 결과의 만족도는 내려가게 된다. 따라서 사용자의 검색엔진에 대한 전체적인 만족도는 검색 품질의 만족도뿐만 아니라 검색 속도도 중요한 요인이다. 따라서 우리는 개인화된 문장 요약 결과를 보여주면서도 검색 속도가 크게 느려지지 않는 문장 요약 시스템 아키텍처를 제안하며, 실제 실험을 통하여 검색 품질과 속도를 만족시킴을 보일 것이다. 또한 개인화 검색은 사용자 프로파일의 저장으로 인한 프라이버시 위협 등의 문제가 따른다. 이러한 문제를 해결하기 위해서 우리는 과거의 연구를 통하여 해결책을 제안하였으며[2], 본 논문을 통해서는 개인화된 문장 요약 시스템과 관련된 시스템 아키텍처만을 제안하고자 하며, 구성은 다음과 같다. 2장에서는 논문의 주제와 관련된 지식들을 설명하며, 3장은 제안하는 시스템 아키텍처를 제시하며, 4장에서는 검색 만족도와 검색 속도에 대한 실험 결과를 보이며, 마지막으로 5장에서 본 논문의 결론과 향후연구 방향을 제시하고자 한다.

2. 관련연구

2.1. 문서 요약

문서 요약의 목적은 문서로부터 추출된 정보를 가지고, 사용자 혹은 어플리케이션의 필요에 따라 가장 중요한 문서의 내용을 요약된 형태로 보여주는 것이며[3], 기존의 연구에서 문서 요약은 추출(Extract)과 요약(Abstract)으로 구분하여, 추출은 원본으로부터 그대로 추출된 부분들로 요약을 구성하는 것이며, 요약은 원본의 내용을 기술하는 새로운 어법들(phrasings)로 구성된 것이라고 정의하였다[4]. 실제 검색엔진에서 사용하는 문서 요약은 사용자에게 해당 웹 문서의 내용을 짧게 보여주는 추출의 방법을 사용하게 되며, 앞으로의 본 논문에서 언급하는 문서 요약은 추출의 방법을 의미한다. 또한 문서 요약을 위해서는 문서 내에서 가장 중요한 문장을 선택하기 위해 문장의 특징을 분석하는 단계가 필요하며, 문장이나 단어의 사전적, 통계적 관련성이나 구의 패턴 매칭등에 의해 중요한 문장을 결정하게 된다. 따라서 중요한 문장을 결정하기 위해 단어의 통계적 정보, 문장의 위치, 문장들 간의 유사도, 문서의 제목 등을 이용하는 다양한 방법들이 연구되어 왔다[3,5,6].

2.2. 개인화 문서 요약

검색엔진에서 개인화 검색에 대한 다양한 연구들이 진행되고 있으며[7,8,9], 사용자가 방문하기 위하여 선택하는 페이지들과 질의어를 전송하는 것과 같은 행동 정보를 통하여 사용자의 특징을 파악하여 얻어지는 정보인 사용자 프로파일을 통하여 질의어를 확장하거나, 결과 문서 순위를 재조정하는 방법을 통하여 개인화된 결과를 보여주게 된다. 문서 요약은 문서에 포함된 다수의 문장 중에서 사용자의 질의어가 포함된 문장 중 문서를 대표하는 문장을 선택하여 보여주게 되며, 이러한 문장의 선택 방법은 정보검색에서 다수의 문서 집합으로부터 질의어와 연관이 높은 문서를 선택하는 것과 동일하게 생각할 수 있다[10]. 따라서 문서 개인화를 위한 다양한 방법들을 동일하게 적용 가능하며, 개인화 문장

요약을 제공하기 위해 사용자의 프로파일을 통하여 각 사용자의 선호도에 적합한 문장을 선택, 추출하기 위한 연구들이 진행되고 있다[3,11].

2.3. 개인화 결과의 평가

일반적인 문서 요약 결과의 평가는 다양한 평가 방법이 존재하지만, 과제(task)의 복잡성으로 인해 아직 가장 최고의 평가 방법은 정해지지 않았다[13]. 또한, 사용자의 선호도에 따라 문서 요약 결과의 만족도가 다른 개인화 결과의 측정은 더욱 복잡해지기 때문에 가장 최고의 평가 방법을 결정하는 것은 더욱 어렵다.

하지만, 앞 절에서 언급한 것처럼 문서 요약은 정보검색의 문서 선택과 동일하게 볼 수 있기 때문에, 정보검색의 다양한 평가 방법을 활용할 수 있게 된다. 따라서 우리는 문서 요약 결과의 만족도를 평가하기 위해 정보검색 분야에서 검색엔진 알고리즘의 효율성을 평가하는 DCG (Discounted Cumulative Gain)를 통하여 제안하는 시스템의 효율성을 평가 하였으며, DCG의 평가 방법은 다음과 같다.

$$DCG_p = r_1 + \sum_{i=2}^{p} \frac{r_i}{\log_2 i}$$

i : 결과 목록에서의 위치(랭크)
r_i : i 번째(랭크) 결과의 연관도 점수
p : 평가하고자 하는 검색 목록의 크기

DCG는 결과 목록의 위치에 기반한 문서(문장)의 유용성 혹은 게인(gain)을 평가하는 것이다. 여기서 게인은 결과 리스트의 최상위부터 마지막 까지 각 게인 값(하위 랭크는 더 낮은 게인 값을 갖는다)을 누적하여 더한 것이다. 예를 들어 사용자 A가 5개의 검색 결과에 <표 1>과 같이 평가를 하였을 경우,

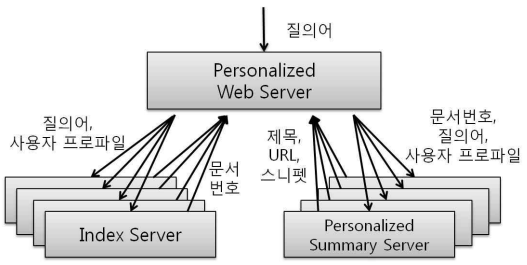
<표 1> 임의의 사용자 평가표

i	r_i	$\log_2 i$	$r_i/\log_2 i$
1	3	N/A	N/A
2	1	1	1
3	2	1.58	1.26
4	1	2	0.5
5	2	2.32	0.86

DCG_5 값은 3+(1+1.26+0.5+0.86) = 6.62가 된다. 따라서 위의 방법에 의해 DCG 값은 결과 목록의 하위보다 상위에 위치한 결과들의 연관도가 높은 경우에 더 높은 DCG 값을 가지며, 높은 DCG 값은 사용자에게 만족스러운 결과를 의미한다.

3. 제안하는 개인화 문서 요약 시스템

검색엔진에서의 문서 요약 시스템은 웹 서버에서 문서 번호(해당 URL을 저장한 문서)와 질의어를 전송하게 되면 타이틀, URL, 요약 정보(snippet)를 웹 서버로 보내 주게 된다[12]. 하지만 개인화 문장 요약 시스템을 위해서는 문서 번호와 질의어에 추가적으로 사용자의 선호도에 적합한 사용자 프로파일 정보를 질의어와 함께 전송해야 하며, [12]를 기반으로 전체적인 검색엔진 시스템 구조와 동작과정에 대해 간략히 알아보면 (그림 1)과 같다.

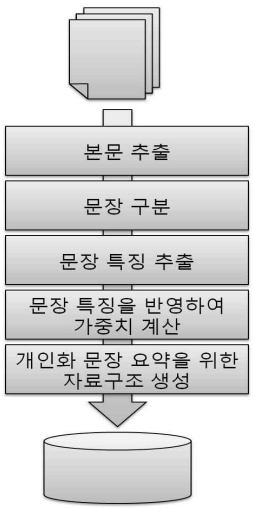


(그림 1) 검색엔진 시스템 구조와 동작과정

사용자가 질의어를 전송하면, 웹 서버는 사용자를 구분하여 사용자 프로파일과 함께 인덱스 서버에 전송하게 된다. 인덱스 서버는 해당 정보를 바탕으로 사용자의 검색 의도와 연관이 높은

문서들을 선택하여 문서 번호를 웹 서버로 넘겨준다. 이렇게 사용자의 프로파일에 따라 랭킹순으로 정렬된 문서들의 요약 정보를 얻기 위해 개인화 요약 서버로 문서의 번호와 질의어, 사용자 프로파일 정보를 전달하여 제목과 URL, 개인화된 문서 요약을 얻게 된다.

따라서 우리가 제안하는 개인화 문서 요약 시스템은 전송받은 정보들을 통해 빠른 시간 안에 사용자의 검색 의도에 적합한 문장들을 선택하여 전달해 주며, (그림 2)와 같은 단계를 통하여 해당 문서에 대한 문장 정보를 저장하게 된다. 각 단계의 방법의 선택은 각 단계와 관련된 다양한 연구들 중, 시스템에 가장 효율이 좋은 방법을 적용하면 최고의 성능을 얻을 수 있을 것이다.



(그림 2) 문서 요약 시스템의 문서 정보 저장 과정

3.1. 본문 추출

문서 내에는 실제 문서의 내용뿐만 아니라 광고나 링크 정보, 스팸(spam) 정보 등의 불필요한 정보들도 많이 포함되어 있다. 이러한 불필요한 정보들은 전체 문서 요약 시스템의 성능 저하와 비효율성을 가져오며, 반대로 올바른 정보들이 제거되는 경우도 성능 저하를 가져오게 된다. 따라서 전체 시스템 성능 향상을 위해 자동으로 해당 문서에서 불필요한 내용들을 제외하고 문서 내의 주요 내용만을 추출하기 위한 다양한 연구들이 진행되고 있으며, 본 논문에서는 정확한 개인화 검색 만족도의 평가를 위하여 수동으로 해당 문서의 주요 내용 부분만을 추출하여 사용하였다.

3.2. 문장 구분

기본적으로 문장 구분은 보통의 경우 .!?를 문장의 끝으로 생각하면 된다. 하지만 .은 소숫점이나 약어인 경우 등의 다양한 문제로 인해 100% 에러 없는 문장 구분은 불가능하며, 문장 구분의 문제역시 하나의 연구 주제로 진행되고 있으며, 앞 절과 마찬가지로 정확한 평가를 위해 수동으로 문장을 구분하였다.

3.3. 문장 특징 추출

문서에 포함된 문장의 중요도를 계산하기 위해 각 문장의 단어의 통계적 정보, 단어의 사전적, 통계적 관련성이나 구의 패턴 매칭 등에 의해 중요한 문장을 결정하게 되며, 이에 관련된 다양한 연구가 진행되고 있다[3,5,6]. 따라서 각 시스템에서 사용되는 문서의 말뭉치(Corpus) 특징에 따라 가장 개인화 만족도가 높은 방법을 사용하면 될 것이다. 본 논문의 실험에서는 간단하게 단어의 빈도수, 고유명사, 문서의 제목, 문장의 위치의 통계적 정보들을 활용하였다.

3.4. 자료 저장

문서에 포함된 각 문장의 가중치 정보를 저장하고, 검색엔진 서버로부터 요청이 왔을 때, 빠른 시간 안에 개인화된 중요한 문장을 전송하기 위하여 <표 2>와 <표 3>과 같이 정보를 저장할 수 있는 자료구조를 사용하였으며, 본 논문의 실험에서는 데이터베이스를 이용하여 구성하였다.

<표 2> 문장 정보의 저장을 위한 자료구조

인덱스 번호	문장 번호	문장 가중치
1	2	0.72
2	10	0.68
3	8	0.65
...	...	...

<표 2>는 각 문장의 가중치 순으로 정렬된 문장 정보 저장을 위한 자료구조이다. 즉 인덱스 번호가 낮은 문장이 더 높은 가중치를 가지며, 문장번호는 인덱스 번호 문장의 문서 내 위치를 나타낸다.

<표 3> 단어 정보의 저장을 위한 자료구조

중요 단어	단어가 포함된 문장 (S)
apple	0000 ... 01101
ipad	0000 ... 01110
ipod	0001 ... 01010
...	...

<표 3>은 문장 내에 포함된 스톱 워드(stop word)를 제외한 알파벳순으로 정렬된 단어들의 저장을 위한 자료구조이다. 본 논문의 실험에서는 문서 내의 최대 문장을 32문장으로 가정하여 단어가 포함된 문장을 나타내는 S값을 32bit int 형을 사용하였다. 32개의 각 비트들은 최하위 비트가 <표 2>의 1번째 인덱스 번호를 최상위 비트는 32번째 인덱스 번호를 가리키며, 해당 단어가 포함된 문장의 인덱스 번호 비트들은 1이 된다. 즉, <표 3>의 apple이란 단어가 포함된 문장은 인덱스 번호 1, 3, 4번의 문장이 된다.

3.5. 개인화 문장 검색

검색엔진의 문서 요약 결과에서 가장 중요한 것은 일반적인 요약보다 사용자의 전송한 질의어가 포함된 문장을 보여주는 것이다[13]. 따라서 다른 문장의 가중치가 높더라도 사용자의 질의어가 포함된 문장이 존재한다면 해당 문장을 먼저 보여주고, 만약 질의어가 포함된 문장이 다수 존재한다면 사용자의 프로파일과 연관도가 높은 문장을 선택하는 것이 사용자 만족도에 더 좋은 결과를 가져올 것이다. 따라서 아래와 같은 방법으로 문장을 선택하여 사용자의 만족도를 향상 시킬 수 있도록 하였다.

사용자 프로파일의 단어가 포함된 문장 확인

1-1) 문장이 존재하는 경우 (개인화 요약)
질의어가 포함된 문장 중, 사용자 프로파일과 연관도가 높은 n개의 문장을 선택

1-2) 문장이 존재하지 않는 경우 (일반 요약)
질의어가 포함된 문장 중,
가장 가중치가 높은 n개의 문장을 선택

※ 질의어가 다수일 경우는, 질의어가 많이 포함된 문장을 먼저 선택

먼저, 일반적인 요약에서는 가중치가 높은 n개의 문장을 선택하기 위해 문서에 포함된 각 문장의 가중치를 구하기 위하여 다음과 같은 공식을 사용한다[3].

$$W_{S_n} = \frac{\alpha A_{s_n} + \beta B_{s_n} \ldots}{\alpha + \beta + \ldots}$$

S_n : 문장 번호

A, B, ... 는 문장의 특징 구분을 위한 방법을 나타내며, $\alpha, \beta, \ldots$ 는 각 방법에 대한 가중치를 나타낸다. 예를 들어, A가 단어의 빈도수를 반영하는 방법이라면 문장의 빈도수를 계산한 가중치 값, B가 문장의 위치를 반영하는 방법이라면 문장의 위치에 따른 가중치 값이 되게 된다. 또한 시스템의 특성에 따라 각 특징에 가중치 $\alpha, \beta, \ldots$ 를 반영하고, 평준화(normailze)하여 최종 가중치 W를 얻게 된다. 이러한 가중치 W를 통해 일반적인 문서 내의 중요도를 반영한 요약 결과를 제공할 수 있지만, 사용자의 선호도에 따른 개인화된 문서 요약 결과는 제공하지 못한다. 따라서 개인화된 문서 요약을 위해 추가적으로 사용자 프로파일 정보를 활용하여 다음과 같은 공식을 통해 가중치를 계산할 수 있다[3].

$$PW_{S_n} = \frac{\alpha W_{s_n} + \beta P_{s_n}}{\alpha + \beta}$$

P_{S_n} : 사용자 프로파일을 반영한 방법

이러한 PW는 사용자의 프로파일 정보를 반영하였기 때문에, PW를 통한 문서 요약 결과의 사용자 만족도는 증가하게 될 것이다. 하지만 사용자의 프로파일은 사용자마다 다르기 때문에 개인화된 문서 요약을 제공하기 위해서는 항상 각 문장마다 사용자 프로파일을 반영하여 가중치를 계산하게 된다. 따라서 사용자 만족도는 증가할 수 있지만, 전체적인 속도는 느려지게 된다. 따라서 우리는 앞 절에서 저장한 데이터를 통해 사용자의 선호도가 높은 문장을 빠른 시간 안에 선택하여 문서 요약을 제공하는 간단한 방법을 제안할 것이며, 다음 장의 실험을 통하여 우리가 제안하는 방법만으로도 충분히 사용자의 만족도를 높일 수 있다는 것을 보일 것이다.

제안하는 시스템에서 개인화된 문장을 선택하는 방법은 다음과 같다.

```
int S[Max_Size]; /* 선택된 문장 번호 저장을 위한 배열 */
N = 전체 질의어 개수;

n = 0;

/* 사용자 프로파일 단어 체크 */
if ( ( P = 사용자 프로파일 단어들의 S값 OR 연산 ) == 0 )
{
    /* 일반적인 문서 요약 */

    S[n++] = 각 질의어의 S 값을 비교하여, 질의어가 가장
             많이 포함된 문장들을 찾은 뒤, 해당되는 문장 중
             최하위 비트(가중치가 가장 높은)의 문장 번호 선택;

    N = N - 선택된 문장을 포함하는 질의어 개수;

    while ( N != 0 )
    {
        T = S[n-1]부터 S[0]까지 OR 연산;

        S[n++] = T값을 가지지 않는 모든 질의어의 S값을
                 비교하여, 질의어가 가장 많이 포함된
                 문장들을 찾은 뒤, 해당되는 문장 중
                 최하위 비트(가중치가 가장 높은)의
                 문장 번호 선택;

        N = N - 선택된 문장을 포함하는 질의어 개수;
    }
}
else
{
    /* 개인화 문서 요약 */

    최하위 비트를 선택할 때, P 값과의 AND 연산을
    통해 최하위 비트의 문장 번호 선택

    만약 AND연산의 값이 0인 경우는 일반적인 요약과 동일
}
```

따라서 다음과 같은 단어 정보를 갖는 문서가 있는 경우

중요 단어	단어가 포함된 문장 (S)
apple	0000 ... 01101
ipad	0000 ... 01110
ipod	0001 ... 01010
tablet	0000 ... 10000
...	...

질의어가 1개일 경우의, apple, ipad, tablet은 각각 질의어를 포함하며 최고의 가중치를 가지는 1번, 2번, 5번 문장이 선택 되며, apple+ipad인 경우는 질의어를 가장 많이 포함하는 3번의 문장이 선택되게 되며, apple+ipad+tablet의 경우는 3번과 5번 두 개의 문장이 선택되게 된다.

만약 ipod라는 사용자 프로파일이 존재한다면, 각 문장을 선택할 때, 사용자 프로파일의 단어를 참조하게 되어 apple의 경우 4번, ipad의 경우는 동일하게 2번의 문장이 선택되며, apple+ipad인 경우는 4번, apple+ipad+tablet인 경우는 4번과 5번 문장이 요약문으로 선택된다.

4. 실험 결과

4.1. 검색 만족도

우리는 앞 장에서 간단한 방법으로 개인화 문장을 추출해 주는 시스템을 제안하였다. 이러한 간단한 방법으로 실제 얼마만큼의 성능을 향상시킬 수 있는지 확인하기 위하여 우리는 다음과 같은 실험을 하였다.

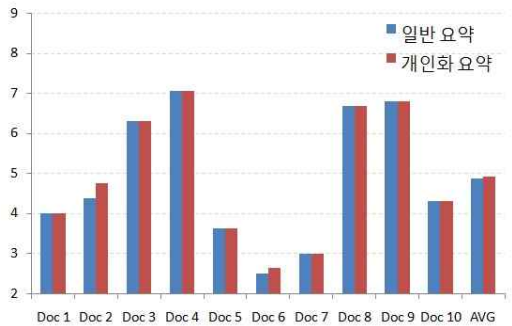
- 임의의 사용자 5명을 선택
- 자신이 검색하고자 하는 정보를 선택
- 정보를 검색하기 위한 질의어를 선택
- 정보를 검색하기 위한 추가 질의어 선택 (3~5)

<표 4> 임의의 사용자 예

검색 의도	각 제조사의 타블렛 PC 비교
질의어	tablet
추가 질의어	apple, samsung, hp

<표 4>와 같은 정보를 바탕으로 상용 검색엔진을 통해 자신의 검색 의도에 적합한 문서(2개의 주제, 주제별 2개의 문서)를 검색하여, 해당 문서내의 질의어가 포함된 각 문장에 대해 자신이 찾는 정보와의 연관도를 0점부터 3점까지 평가하도록 하였다. (점수는 0점에 가까울수록 찾고자 하는 정보와 연관이 적은 문장이며, 3점에 가까울수록 연관이 높은 문장을 뜻한다)

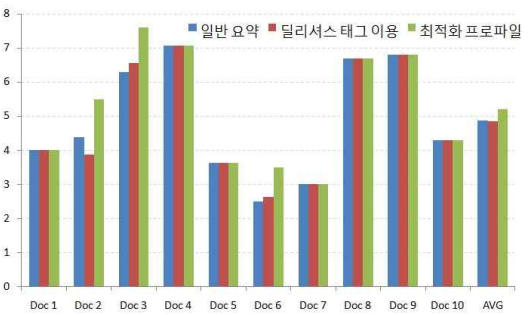
이러한 평가 결과를 바탕으로 임의의 10개 문서에 대해 일반적인 요약 결과와 우리가 제안하는 개인화 요약 방법의 DCG 값을 비교하였으며, 결과는 (그림 3)과 같다.



(그림 3) 일반 요약과 개인화 요약의
DCG 값 비교

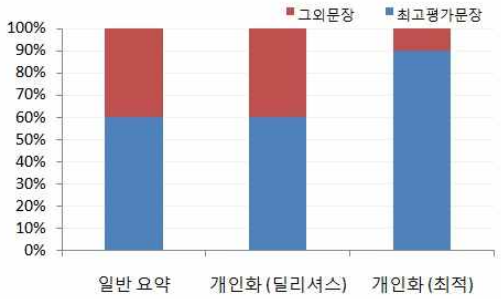
앞 장에서 보았던 방법에서 알 수 있듯이, 사용자의 프로파일의 단어들이 해당 문서에 포함되어 있지 않는 경우나 질의어가 포함된 문장들이 모두 사용자 프로파일의 단어를 포함하는 경우에는 DCG 값에 전혀 영향을 주지 않았으며, 질의어가 포함된 문장들의 평가 점수가 모두 높은 경우에도 DCG 값은 변화가 없었다. 하지만 사용자가 전송한 질의어를 포함한 문장이 많이 존재하고 사용자 프로파일의 단어들이 사용자의 검색 의도에 일치하는 문장들에 많이 집중되어 있는 경우 DCG 값이 증가하였다.

또한 위 실험에서는 동일한 사용자가 직접 질의어와 연관된 단어를 선택하고 각 문장에 대해 평가하여 DCG 값이 하락하는 경우가 없었지만, 잘못된 프로파일로 인하여 오히려 DCG 값이 하락하는 경우도 발생할 수 있기 때문에 사용자의 만족도를 크게 향상시키기 위해서는 사용자의 의도에 맞는 정교한 프로파일의 생성과 활용이 필요할 것이다. 따라서 사용자 프로파일이 사용자의 만족도에 미치는 영향을 확인하기 위해 일반적인 사용자 프로파일과 사용자의 의도에 가장 적합한 이상적인 프로파일을 생성하여 비교하였다. 일반적인 사용자 프로파일로는 사용자가 직접 선택한 단어들이 아닌 딜리셔스 사이트[14]의 연관된 태그 정보를 활용해 질의어로 선택한 태그와 연관도가 높은 3개의 태그를 선택하여 사용자 프로파일로 생성하였다. 또한 사용자의 검색 의도에 가장 이상적인 사용자 프로파일을 생성하기 위해 동일한 주제 내의 모든 문장 중, 사용자의 평가가 높은 문장에서만 자주 나타나는 단어들을 선택하여 사용자 프로파일을 생성하여 비교하였으며, 결과는 (그림 4)와 같다.



(그림 4) 임의의 프로파일과
이상적인 프로파일의 DCG 값 비교

위 실험을 통해 사용자의 프로파일이 사용자의 선호도에 따라 적합하게 선택된다면, 간단한 방법만으로도 사용자 만족도 증가에 영향을 주게 됨을 확인할 수 있었다. 또한 위의 결과에서 사용자 프로파일로 인하여 요약 결과가 변경되었지만, 동일한 DCG 값을 가지는 문서들의 사용자 만족도를 자세히 평가하기 위해 질의어가 포함된 문장들 중, 실험자가 가장 높은 점수를 평가한 문장이 요약 결과에 나타나는지의 여부를 일반적인 문서요약, 딜리셔스 사이트의 태그 정보를 활용한 프로파일, 최적화된 사용자 프로파일을 사용한 경우를 각각 비교하였다.



(그림 5) 문서 요약 결과 비교

위의 결과를 통해 일반적인 문서 요약 결과보다 사용자의 선호도에 적합하게 생성된 사용자 프로파일을 활용한 개인화 문서 요약의 검색결과가 더 높은 만족도를 제공할 수 있음을 확인할 수 있었다. 본 실험의 데이터 집합은 질의어가 포함된 문장 중 가장 가중치가 높은 문장에 대해 사용자가 최고의 점수로 평가한 경우가 많아 일반적인 사용자 프로파일을 사용한 경우와 큰 차이가 없었지만, 최적화된 사용자 프로파일을 가정한 경우 사용자의 만족도가 증가됨을 확인 할 수 있었기 때문에 다 수의 사용자와 다수의 데이터 집합을 사용한다면 좀 더 큰 차이를 발견할 수 있었을 것이다. 따라서 개인화 검색엔진 시스템에서 해당 사용자의 의도에 적합한 문서가 선택될 때, 적합한 사용자 프로파일의 생성과 간단한 방법만으로 충분히 사용자의 만족도를 향상 시킬 수 있는 개인화된 결과를 보여줄 수 있음을 확인할 수 있었다.

4.2. 검색 속도

상용 검색 엔진인 구글의 경우, 검색 결과에 표시되는 시간을 통해 검색 속도를 쉽게 알 수 있으며 일반적으로 동일한 내용을 검색하게 되면 처음 보다 검색 속도가 빨라지는 것을 확인할 수 있다. 그 이유는 구글의 검색 결과는 사용자에 관계없이 질의어를 전송하여 얻게 되는 문서 랭킹과 요약 정보는 항상 동일하므로 한번 검색된 질의어에 대해 캐쉬 방식으로 검색 서버에 질의어와 문서의 랭킹 결과, 해당 문서에 대한 요약 정보를 저장하여 검색 서버에서 인덱스 서버와 문서 요약 서버(도큐먼트 서버)로의 전송 없이 빠른 시간 안에 문서 요약 정보를 보여주는 방법을 사용하기 때문이다. 하지만 검색 결과에 표시되는 시간은 질의어와 연관이 높은 문서의 랭킹과정의 시간과 문서에 대한 요약 정보를 얻는 시간의 합이기 때문에 결과 시간만을 가지고는 문서 요약 정보만을 얻는데 걸린 시간만을 정확히 알기 어려우며, 또한 구글에서 사용하는 문서 요약 서버의 성능과 처리 방법등에 대한 정확한 정보가 없기 때문에 제안하는 시스템과의 정확한 비교가 어렵다. 따라서 제안하는 문서 요약 시스템의 검색 속도 결과를 통해 전체 검색 시간에 미치는 영향을 예상해 보았다.

제안하는 시스템 및 모든 개인화 검색 시스템은 개인화 검색 결과의 제공으로 인해 각 문서에 동일한 질의어를 전송하여도 각 사용자 프로파일에 따라 다른 결과를 얻게 된다. 따라서 캐쉬 방식을 통한 속도의 향상을 얻을 수 없고, 각 검색마다 사용자 프로파일을 반영한 개인화 요약이 진행되어 전체적인 검색 속도의 하락은 불가피하게 된다. 하지만, INTEL XEON QUAD CORE CPU 2개, 32GB RAM, Linux 운영체제에서 Postgres 데이터베이스를 사용하여 우리가 제안하는 방법을 통해 개인화된 요약 결과를 제공하는 서버에 질의어와 사용자 프로파일을 전송한 경우 1개의 문서(약 20문장)를 처리하는데, 약 0.4초 내외의 시간이 걸렸다. 따라서 분산으로 전체 개인화 검색 엔진 시스템을 구성하고 처리하는 경우 검색 시간의 큰 증가는 없을 것으로 예상된다.

5. 결론 및 향후연구

우리는 간단한 방법을 통해서 검색 시간이 크게 느려지지 않으면서도 사용자의 만족도를 향상 시킬 수 있는 개인화 문서 요약 시스템을 제안하였다. 또한 임의의 데이터를 사용한 실험을 통하여 사용자의 만족도가 향상됨을 보였다. 하지만, 실제 검색엔진 시스템에 적용하기 위해서는 정확한 사용자 프로파일의 생성과 어떠한 정보를 통해 프로파일을 생성할 것인지와 잘못된 프로파일의 생성으로 인하여 오히려 혼란을 가져오는 경우의 해결에 대한 연구도 필요할 것이다. 따라서 전체 검색엔진 시스템을 구성하여, 실제 사용자 프로파일을 생성하고 개인화 문서 결과를 제공하는 개인화 검색엔진 시스템에서 우리가 제안하는 시스템을 적용하여, 전체적인 시스템의 성능과 각 단계의 문제점을 분석 보완하는 단계가 필요할 것이다. 따라서 향후 연구를 통해 이에 대한 시스템을 구현하여 개선 및 발전시키는 연구를 진행해 나갈 것이다.

참 고 문 헌

[1] I. Varlamis and S. Stamou, “Semantically driven snippet selection for supporting focused web searches”, Data&Knowledge Engineering, 2008.

[2] 김동욱, 강수용, 김한준, 이병정 “폭소노미 기반 개인화 웹 검색 시스템”, 디지털콘텐츠학회 논문지 제11권 제1호 (2010. 3)

[3] Diaz and Gervás, 2007 A. Diaz and P. Gervás, User-model based personalized summarization, Information Processing & Management 43 (6) (2007), pp. 1715 - 1734.

[4] Hovy, E.H. and C-Y. Lin. 1998. Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press.

[5] Hahn, U., & Mani, I. (2000). The Challenges of Automatic Summarization. Computer, 33(11), 29-36

[6] Mani, I. & Bloedorn, I. (1998). Machine learning of generic and user-focused summarization. In Proceedings of the 15th national conference on artificial intelligence, Menlon Park, California, (pp. 821-826)

[7] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In Proceedings of CIKM '05, pages 824-831, 2005.

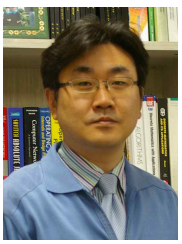
[8] Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch S.: Personalized Search on the World Wide Web. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2007) this volume

[9] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring

Folksonomy for Personalized Search. In Proc. of SIGIR' 08, 2008, 155-162.
[10] Goldstein, J., Kantrowitz, M., Mittal, V., and Carbon ell, J. 1999. Summarizing Text Documents: Sentence Sel ection and Evaluation Metrics. SIGIR'99.
[11] Mana, M., Buenaga, M. & Gomez, J. M. (1999). Usin g and evaluating user directed summaries to improve information access. In Proceedings of the third euopean conference on research and advanced technology for digi tal libraries (pp. 198-214).
[12] Barroso, L. A., Dean, J., and Urs Hölzle, U. 2003. Web search for a planet: The Google cluster architectur e. IEEE Micro 23, 2, 22-28.
[13] Mana, M., Buenaga, M., & Go´mez, J. M. (1999). Using and evaluating user directed summaries to impro ve information access. In Proceedings of the third europe an conference on research and advanced technology for digi tal libraries (pp. 198 - 214). LNCS 1696: Springer-Verlag.
[14] http://delicious.org/

김 동 욱

2009년 : 호서대학교 컴퓨터공학부 (학사)
현 재 : 한양대학교 전자컴퓨터통신공학과 (석사)
관심분야 : 정보 검색(Information Retrieval), 파일 시스템, 플래시메모리 기반 저장 시스템



강 수 용

1996년 : 서울대학교 수학과(학사)
1998년 : 서울대학교 전산과학과 (석사)
2002년 : 서울대학교 전기컴퓨터공학부 (박사)
2003년～현 재: 한양대학교 컴퓨터공학부 교수
관심분야 : 멀티미디어 시스템, 분산시스템, 플래시 메모리 기반 저장 시스템 등

김 한 준

1994년 : 서울대학교 계산통계학과 졸업 (이학사)
1996년 : 서울대학교 전산과학과 대학원 졸업 (이학석사)
2002년 : 서울대학교 컴퓨터공학부 대학원 졸업 (공학박사)
2002년～2002년 : 서울대학교 공과대학 박사후 연수 과정
2002년～현 재 : 서울시립대학교 전자전기컴퓨터공학부 교수
관심분야 : 정보검색(Information Retrieval), 기계학습(Machine Learning), 데이터마이닝(Data Mining), 데이터베이스(Databases) 등

이 병 정

1990 : 서울대학교 계산통계학 (학사)
1998 : 서울대학교 전산과학(석사)
2002 : 서울대학교 컴퓨터공학 (박사)
1990년～1998년: 현대 전자 SW연구소
2002년～현 재: 서울시립대학교 컴퓨터과학부 교수
관심분야 : 소프트웨어 진화, 개발 방법론, 소프트웨어 품질 등

장 재 영

1992 : 서울대학교 계산통계학과 (학사)
1994 : 서울대학교 계산통계학과 전산과학전공 대학원(석사)
1999 : 서울대학교 계산통계학과 전산과학전공 대학원(박사)
2000년～현 재: 한성대학교 컴퓨터공학과 교수
관심분야 : 데이터베이스, 데이터마이닝