개인 기호정보 필터링을 사용한 모바일 시맨틱 검색

전호철°, 김태환, 최중민
한양대학교 컴퓨터공학과
e-mail : hcjeon@cse.hanyang.ac.kr, kimth@cse.hanyang.ac.kr, jmchoi@hanyang.ac.kr

Mobile Semantic Search using Personal Preference Filtering

Ho-Chul Jeon°, Tae-Hwan Kim, Joong-Min Choi
Dept. of Computer Science & Engineering, Han-Yang University

요 약

본 논문에서는 모바일폰을 통한 시맨틱 검색 및 개인 기호정보를 사용한 검색 결과의 필터링이 가능한 시스템을 제안 하고자 한다. 시스템에서는 모바일 컨텐츠와 웹 컨텐츠의 검색 연동, 사용자 기호정보의 유·무선 장치의 공유 및 유·무선 장치간 검색 결과의 공유를 지원한다. 모바일폰의 컴퓨팅 능력을 고려해 모바일폰에는 사용자 인터페이스만을 유지 하도록 한다. 모바일폰을 통한 시맨틱 검색을 지원하기 위해 시스템은 실험적으로 뉴스 도메인에 국한된 카테고리에 대한 분류 체계 온톨로지를 구축하며, 각 카테고리간 관계를 설정 한다. 또한, 개인 기호정보를 통한 검색 결과의 필터링을 위해 사용자 기호정보를 XML 형태의 벡터 모델로 유지하며, 이는 서버의 데이터베이스에 각 사용자 계정으로 저장하고 공유한다. 모바일폰의 여러 단점을 극복하고 장점을 극대화 하기 위해 검색 결과를 서버에 저장하고 이를 유·무선 장치간 상호 공유 할 수 있도록 한다. 본 논문에서는 시스템의 아키텍처와 구성 및 주요 기능에 대해서 기술하고자 한다.

1. 서론

모바일폰의 폭발적인 보급과 다양한 사용자의 요구에 따라 모바일폰 검색 요구가 증가하고 있고, 최근 이를 반영한 다양한 모바일폰들이 보급되고 있다. 많은 모바일 장치들이 개발되고 보급되고 있지만 모바일 장치들은 태생적인 한계를 지니고 있다. 전원의 한계, 디스플레이 화면의 한계, 전송 능력(bandwidth)의 한계, 처리능력(computation power)의 한계, 단절(disconnection) 등이 그것이다. 반면에 모바일 장치의 가장 큰 장점은 이동성과 편의성이다. 이는 곧 어디서나 사용 가능 하다는 것을 의미한다.

모바일폰을 통한 검색을 지원하면서 이러한 장치들이 지니고 있는 근본적인 문제점들을 해결하기 위해 시맨틱 검색이 적합한 방법 중 하나다. 아울러, 보다 효과적이고 효율적인 검색을 지원하기 위해 개인별 선호정보를 사용해서 검색 결과에 대한 필터링을 사용한다. 이를 통해 보다 개인화된 검색결과를 얻을 수 있으며, 사용자는 많은 양의 불필요한 데이터를 접하는 기회가 적어지게 된다.

모바일 장치의 한계를 극복하기 위해 시스템에서 제공하는 또 다른 기능은 사용자 기호정보의 공유와 검색 결과의 공유다.

모바일 장치를 통한 검색을 할 때, 지속적으로 서버와의 연결을 유지 하지 않고 검색 결과를 서버에 저장하고, 이후에 사용함으로써 모바일 장치의 가장 큰 단점인 단절을 어느 정도 해결 할 수 있으며, 검색 결과가 매우 많은 경우 처리능력과 전송능력의 한계로 인한 검색 수행 시간이 길어지는 문제를 해결할 수 있다.

또한, PC 에서 검색한 결과를 서버에 저장하고, 이를 모바일폰에서 공유하도록 함으로써 장치에 상관없이 어디서나 검색 결과를 재확인 할 수 있다.

사용자 기호정보를 공유 함으로써, 사용자는 사용 장치에 상관없이 개인화된 정보 및 개인별 필터링된 최적의 검색 결과를 얻을 수 있다.

본 논문에서는 앞서 기술한 여러 기능들을 사용함으로써 모바일 장치의 최대 장점인 이동성과 편의성을 통해 언제 어디서나, 사용자에게 최적화된 검색결과를 제공 할 수 있는 시스템을 제안 하고자 한다.

본 논문의 구성은 다음과 같다.
먼저 2 장에서는 관련된 연구를 통해서 모바일폰 검색, 모바일 장치에서의 기호정보 활용에 대해 알아보고, 3 장에서는 시스템 구성에 대해 살펴본다. 4 장에서는 각 기능에 대한 정책 및 방법에 대해 언급하고 마지막으로 5 장에서 결론 및 향후 연구 과제에 대해 논한다.

2. 관련연구

이 장에서는 모바일폰 검색 및 모바일폰에서 사용하는 개인 기호정보에 대한 관련 연구를 살펴보고, 기존 시스템의 문제점들을 살펴 보도록 한다.

2.1 모바일폰에서의 기호도(Preference)

모바일 장치에서 개인의 기호 정보를 활용하는 다

양한 연구들이 진행 중이다.

Mobile Electronic Personality[1]는 사용자의 기호가 상황과 장소에 따라 변하는 것에 초점을 맞추어 상황정보인지 후 적합한 기호정보를 선택하고 적용한다. 기호정보는 모바일폰에 저장되며 모바일폰에서 상황 정보를 인지한다.

PANS[2]는 사용자의 기호 정보에 따라 상점에 있는 스크린을 통해 사용자가 선호하는 스포츠의 경기결과를 제공한다. 이 시스템에서는 사용자의 기호 정보를 사용자의 모바일 장치에 저장하고, 모바일 장치에 있는 기호정보 제공자(Preference Provider)와 스크린인 서비스 제공자(Service Provider), 사용자의 기호도를 수집하는 기호정보 요구자(Preference Requester)간 상호작용을 통해 스크린에 사용자 기호에 적합한 스포츠 경기결과를 제공한다. 기술한 방식의 기호도 정보 수집 및 이에 기반한 정보 제공은 개인과 어떠한 상호작용 없이 이루어 지며, 기호정보가 일정 기간을 두고 변화하는 사용자의 기호를 반영하지 못하는 단점이 있다. 또한 이러한 시스템들은 기호 정보를 모바일 장치에 저장 하도록 함으로써 모바일 장치의 한계를 극복하고 효율적으로 활용하기가 어렵다.

2.2 모바일폰 검색

모바일 검색 서비스는 점점 중요한 사용자 행동이 돼가고 있고, 구글, 야후, 마이크로소프트 등 기존 모바일 서치 엔진들은 모바일 검색을 위한 적절한 서비스를 제공하기에 적합하지 않다. 대부분의 질의기반 검색과 리스트 기반 결과 표현 형식은 모바일 장치의 입출력 특성에 최적화되지 않았다. 예를 들면 간단하게 웹 컨텐츠를 모바일 장치에서 표현 가능한 형태로 번역 하는 정도다[3].

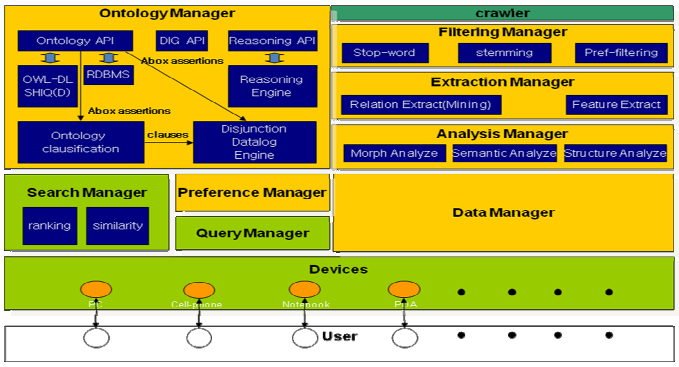
[3]에서는 모바일 장치의 화면 출력 특성에 보다 적합한 결과 출력 방법을 제안하고 현존하는 7 개의 모바일 검색 엔진을 비교 평가한다. 검색 결과를 표시 할 때, 문서의 일부분을 나타내는 대신 과거 해당 검색 결과를 선택했던 질의 단어들을 표시 해줌으로써 제한된 모바일 화면 공간을 절약 할 수 있고, 관련된 정보들을 쉽게 알 수 있다.

[4]에서는 이전의 다른 사람이 수행한 질의어와 위치 정보를 고려해 모바일폰을 통한 검색시 이를 활용한다. 이것은 동일한 위치에서 행해진 다른 사람들의 질의어가 현재의 사용자 질어에도 영향을 미치는 것을 실험을 통해 설명하고 있다.

[3],[4] 모두 검색 결과를 저장 하거나 재사용 할 수 없고, 정보 검색시 질의어에 대한 사용자의 의도를 파악하기 어렵다는 단점이 있으며, 이 역시 모바일 장치의 한계를 극복하고 효율적으로 검색 하기는 어렵다.

3. 시스템

3.1 시스템 구성



(그림 1) 모바일 시맨틱 검색 시스템 구조

본 논문에서 제안하는 모바일 시맨틱 검색 시스템은 온톨로지 매니저, 기호정보 매니저, 필터링 매니저, 추출 매니저, 분석 매니저, 데이터 매니저, 쿼리 매니저, 검색 매니저, 크롤러로 구성된다.

사용자는 다양한 장치들을 사용해서 검색 매니저와 상호작용해서 검색 결과를 얻을 수 있으며, 데이터 매니저를 통해 검색 결과를 저장하거나 저장된 검색 결과를 로드할 수 있다.

온톨로지 매니저는 온톨로지 구성과 시맨틱 검색, 관련결과 추론 등의 기능을 한다. 기호정보 매니저는 사용자 개인별 기호정보의 생성, 수정, 삭제를 위한 관리 기능을 한다. 또한 기호정보 관리자는 기호정보 공유를 위해 데이터베이스와 상호작용한다. 필터링 매니저는 기호정보 매니저와 상호작용하며, 이러한 정보를 활용해 보다 개인에게 적합한 결과를 추출해 낸다. 추출 매니저는 온토로지 구성 시 뉴스 카테고리의 분류 체계에 대한 정보 및 각 분류별 관계들을 추출하고 관리한다. 분석 매니저는 크롤러를 통해 수집된 문서들을 색인하는 기능을 하며, 이때, 형태소 분석과 의미, 구문 분석을 수행한다. 데이터 매니저는 기호정보, 검색 결과를 공유하기 위해 데이터베이스에 저장하거나 데이터베이스에서 저장된 정보를 로드 할 때 각 접속 장치에 따른 적합한 형태로 변환하는 기능을 한다. 검색 매니저와 쿼리 매니저는 사용자의 접속 장치와 직접 상호작용하며, 쿼리 매니저는 사용자의 질의어에 대한 전처리를 수행한다. 검색 매니저는 사용자 질의어에 대한 유사도 측정 및 검색 결과의 랭킹을 수행한다.

3.2 온톨로지 구성

온톨로지는 SWRL 을 이용해 보다 정확한 의미검색을 제공하고자 하며, 이는 사용자의 질의에 가장 적합한 결과를 얻기 위한 사용자 규칙을 사용하는 것을 의미한다.

온톨로지는 시맨틱 검색을 하기 위해 기본적으로 구성해야 하며, 본 논문에서는 뉴스를 대상으로 구성했다.

각 뉴스가 속하는 카테고리별 분류체계는 수동으로 작성되며, 추론 시 TBox 로 사용된다. 각 카테고리별 분류 체계에는 해당 키워드들을 적용하며, 이는 각 카테고리에 대한 특징으로 규정한다.

카테고리간 관계의 규정 역시 수동으로 작성되며, 이는 관련 뉴스들을 찾을 때 추론을 위해 사용된다.

크롤러를 통해 자동으로 수집된 뉴스들은 각 카테고리별 지정된 키워드에 따라 분류된다. 이때, 자동으로 분류 되지 못한 뉴스들은 관리자에 의해 수동으로 분류된다.

3.3 유사도 측정 및 시맨틱 정보를 활용한 관련 결과 추론

사용자 질의어에 대해 각 카테고리별 특징 집합에 대한 유사도를 측정하며, 최상위 카테고리내 모든 뉴스들에 대해 유사도를 측정한다.

이때, 유사도 측정은 정보검색 시스템에서 일반적으로 사용하는 코사인유사도법을 사용한다.

$$sim(D,Q) = \frac{D \cdot Q}{|D||Q|} = \frac{1}{W_d W_q} \sum_t w_{dt} \cdot w_{qt} \quad (1)$$

$$\Rightarrow sim(D,Q) = \frac{1}{W_d W_q} \sum_{t \in Q} tf_{dt} \cdot \left(\log \frac{N}{df_t} \right)^2 \quad (2)$$

식 (2)를 통해 계산된 유사도를 사용해 상위 n 개의 문서를 추출하고, 이 n 개의 뉴스들에 대해 추론을 적용한다. n 은 실험을 통해 3 으로 사용한다.

4. 시스템 기능

본 장에서는 시스템에서 제공하는 각 기능에 대해 보다 구체적으로 기술하며, 각 기능을 통해 어떻게 모바일 장치의 한계를 극복하고, 어떻게 모바일 장치의 장점을 극대화하는지를 나타내고, 각 기능을 위해 상호작용하는 시스템 요소들에 대해 기술한다.

4.1 웹 컨텐츠 및 모바일 컨텐츠의 검색

시스템에서는 모바일 컨텐츠에 대한 검색 뿐만 웹 컨텐츠의 검색을 함께 제공한다. 이는 아직까지 모바일 컨텐츠가 많이 활성화 되지 않았고, 이는 빈약한 검색 결과를 초래하게 된다. 웹 컨텐츠의 검색은 시스템에 의해 자동으로 행해지며, 모바일 컨텐츠의 검색 결과가 없거나 일정 임계값(α) 이하의 경우 메타검색을 통해 웹 컨텐츠를 검색한다. 이렇게 검색된 웹 컨텐츠는 사용자에게 결과로 전달 된 뒤, 시스템에 저장된다.

4.2 검색 결과의 저장 및 공유

검색 결과를 저장함으로써 사용자는 언제든지 장치에 상관없이 검색 결과를 재사용할 수 있다. 이것은 모바일 장치의 여러 한계를 극복할 수 있는 좋은 해결책일 것이다. 예를 들면, 지속적인 서버와의 연결 없이 결과를 서버에 저장하고 나중에 결과를 확인 함으로써 전원의 한계와 전송 능력의 한계, 처리능력의 한계, 단절 등의 한계를 극복할 수 있다. 이러한 시도는 모바일 에이전트[6]의 기본 개념과 같다.

시스템의 크롤러를 통해 자동으로 수집된 정보들은 각 카테고리별 지정된 키워드에 따라 분류되어 저장이 되는데, 여기에는 사용자의 기호정보와, 검색의 결과 정보가 저장이 된다. 그리고 이것은 데이터베이스에 저장된 정보를 각 접속 장치에 따른 적합한 형태로 변환하여 모바일 컨텐츠와 웹 컨텐츠의 공유를 목적으로 한다.

4.3 사용자 기호정보의 공유

사용자 기호정보는 사용자가 선호하는 정보를 가지고 있는 것으로 확장성과 접근 용이성으로 XML 형태로 데이터베이스에 저장된다. 이러한 기호정보의 공유를 통해 모바일 장치는 새로운 사용자 기호정보를 생성하는데 필요한 전력, 메모리, 데이터 전송 등을 절약할 수 있고, 이는 모바일 장치를 통해 개인 기호정보를 관리 하는데 있어 전원의 한계와 전송 능력의 한계, 처리능력의 한계, 단절등의 한계를 극복 가능하게 한다. 사용자 기호정보는 일반적으로 사이트 가입 시 주로 작성되며, 대부분은 웹을 사용할 때 적용되거나 웹을 통한 추천을 위해 사용된다.

그러나 이러한 정적인 사용자 기호정보는 기호가 지속적으로 변화하는 사용자에 대해 부적합하다. 특히 연령층이 낮을수록 기호의 변화는 더 빠르고 다양하다.

기호정보 매니저를 통해 가공된 사용자 개인별 기호정보가 저장되어 있는 데이터베이스와 상호작용한다.

4.4 기호정보 갱신 정책

제안 시스템에서 사용자 기호정보의 갱신은 매 질의시 마다 이루어 지며, 사용자 기호정보의 갱신에 대한 2 가지 정책을 소개한다. 첫째, 사용자 기호정보가 지속적으로 갱신되는 동적 환경에 대한 갱신 정책이 있다. 둘째, 사용자 기호 정보가 사용자의 초기 설정으로 결정되고 이후, 사용자의 의한 수동적인 갱신이 가능한 정적 환경에 대한 정책이 있다.

4.4.1 동적 환경

동적 환경은 사용자 기호정보가 지속적으로 갱신되는 환경을 의미하며, 이때 사용자의 기호정보는 지속적으로 변화한다. 이를 위해 본 논문에서는 다음과 같은 가정을 필요로 한다. “사용자는 자신이 좋아하는 것에 대해 지속적으로 관심을 나타내고 이는 관련 문서의 접근으로 나타난다.”

사용자의 기호 정보에 적용되는 term 들은 3 곳에서 추출될 수 있다. 사용자가 질의하는 질의어에 포함된 term, 카테고리에 포함된 키워드, 사용자가 선택한 검색 결과의 키워드다.

질의어에 포함됐거나 카테고리에 포함된 키워드 또는 사용자가 선택한 검색 결과의 키워드 중에서 df 값이 가장 작은 n 개의 term 이 사용자 기호 정보에 포함된다. 반면에, 사용자 기호 정보에 포함된 각 term 의 가중치값이 일정 임계값 미만인 경우 사용자 기호정보에서 삭제 된다.

$$tf \cdot df \quad (3)$$

각 term 은 TF, DF 값을 유지 하며, 사용자가 매 문서 접근 때 마다 기호정보에 term 이 존재 하는 경우 해당 term 의 TF 와 DF 값은 1 씩 증가하고, 존재 하지 않는 경우 일정 비율(1%)로 감소 한다.

각 term 에 대한 가중치 값은 식 (3)과 같이 TF*DF 로 계산한다.

일반적으로 정보검색 시스템에서는 TF*IDF 를 사용하지만 본 논문에서는 가중치 계산에 IDF 대신 DF 를 사용한다. IDF 는 전체 문서 집합에서 일부 문서에 집중적으로 나타나는 단어에 대해 높은 가중치를 할당하는 방법이다. 반면에, DF 는 전체 문서 집합에서 해당 단어가 나타나는 빈도수를 의미한다. 즉, 보다 많은 문서에서 단어가 나타날수록 높은 가중치를 할당하는 방법이다.

제안 하는 시스템에서 기호도를 특징 지을 수 있는 단어들을 추출하기 위해서는 각 단어가 사용자가 접근하는 여러 문서에서 두루 나타날수록 보다 효과적일 것이다. 때문에 IDF 가 아닌 DF 가 적합한 가중치 계산을 위한 한 요소가 된다.

$$< t_i, tf_i, df_i, w_i > \quad (4)$$

식 (4)는 각 term 에 대한 attribute 들을 포함하는 벡터를 나타낸다. t_i 는 i 번째 term, tf_i 는 i 번째 term t 가 문서에서 나타난 빈도수, df_i 는 term t 가 나타난 문서의 빈도수, w_i 는 term t 의 가중치 값을 의미한다.

4.4.2 정적 환경

정적 환경은 기존 대부분의 웹 사이트에서 제공하는 일반적인 방법으로, 사용자가 초기 가입 시 선호하는 카테고리를 선택한다. 기호정보의 변경은 사용자의 해서만 변경되며, 각 분야에 대해 가중치를 줄 수 없다는 단점이 있다. 반면에, 이는 구현이 쉽고 사용자에 의해 관리 되기 때문에 기호정보는 매우 정확한 장점이 있다.

기호정보에 사용되는 term 들은 카테고리에 지정된 키워드에 국한되며, 사이트 초기 가입 시 사용자가 각 카테고리를 선택함에 따라 해당 키워드들이 기호정보에 포함된다. 삭제 역시 사용자에 의해 수동으로 이루어 진다.

4.5 변환

변환은 검색 결과의 저장을 위해 사용하며, 검색결과를 서버에 저장할 때는 HTML 이나 WML 에서 XML 파일로 변환하며, 이와 반대로 검색 결과를 서버에서 로드 하고자 할 때는 XML 을 WML 이나 HTML 형태로 변환한다. 이때, 검색결과를 요청한 장치가 모바일폰인 경우는 WML 형태로 변환하며, PC 인 경우는 HTML 형태로 변환한다.

4.5.1 WML(Wireless Markup Language)

서버에 저장돼 있는 검색 결과를 모바일폰에서 로드하고자 할 때, XML 형태의 검색결과를 모바일폰에서 표현 가능한 형태인 WML 형태로 변환 한다.

Apache Cocoon[1] 을 사용해 구현 할 것이다.[1]

4.5.2 HTML

서버에 저장된 검색 결과를 PC 에서 로드 하고자 할 때, XML 형태의 검색결과를 HTML 형태로 변환 한다. 오픈 소스인 Apache Cocoon[1] 을 통해 구현 할 것이다.

5. 결론 및 향후 연구

본 논문에서는 모바일폰을 통한 시맨틱 검색 및 개인 기호정보를 사용한 검색 결과의 필터링이 가능한 시스템 아키텍처에 대해 제안했다.

시스템은 또한 단절, 컴퓨팅 능력의 한계, 디스플레이 화면의 제한 등 모바일 장치의 한계를 극복하고 이동성과 편의성을 극대화 하기 위해 사용자 기호정보 공유, 검색 결과의 저장과 공유, 웹 컨텐츠와 모바일 컨텐츠의 검색을 지원하도록 했다.

향후 연구해야 할 것은 용량이 매우 큰 웹 컨테츠를 모바일 장치로 표현하기 어렵기 때문에 보다 효과적인 표현 방법이 필요하고, 기호정보를 활용한 다양한 추천 과 기호정보 기반의 소셜 네트워크 구성을 통한 협업 필터링으로 보다 다양하고 정제된 결과를 제공 할 수 있도록 하는 것이다.

참고문헌

[1] Francesco Ricci and Quang Nhat Nguyen, “Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System”, IEEE Intelligent Systems, 2007

[2] Pekka Jäppinen and Jari Porras, “PANS-Preference-aware news screen”, Proceeding of the seventh IEEE workshop on Mobile Computing Systems & Applications, 2006

[3] Church, K., Smyth, B., and Keane, M. T., “Evaluating Interfaces for Intelligent Mobile Search”, In Proceedings of the 2006 international Cross-Disciplinary Workshop on Web Accessibility (W4A): Building the Mobile Web: Rediscovering Accessibility? (Edinburgh, U.K., May 22 - 22, 2006). W4A, vol. 134. pages 69-78, ACM Press, New York, NY., 2006

[4] Jones, M., Buchanan, G., Harper, R., and Xech, P., “Questions not answers: a novel mobile search technique”, Proceedings of the SIGCHI, 2007

[5] Church, K. and Smyth, B. “Mobile Content Enrichment”, Proceedings of the 2007 International Conference on Intelligent User Interfaces (IUI). In press, 2007

[6] James E. White. “Mobile Agent”, pages 437-472, In Bradshaw, 1996

1 http://cocoon.apache.org/mirror.cgi