시멘틱 환경에서의 개인화 검색

김제민° 박영택
숭실대학교 컴퓨터학과
d5f4g3h2@hanmail.net, Park@computing.ssu.ac.kr

Personalized Search Service in Semantic Web

Je-Min Kim° Young-Tack Park
Dept. of Computer Science, Soongsil University

요 약

웹에 분산된 모든 웹 페이지는 구조가 서로 다르다. 시멘틱 웹 환경은 이형적인 구조를 갖는 웹 페이지들의 메타데이터를 바탕으로 시멘틱 검색이 가능하다. 그러나 일반적으로 사용자의 요구에 따른 시멘틱 검색은 상황에 따라 엄청난 수의 검색 결과를 내놓는다. 따라서 검색 결과에 대해 각 사용자에 맞는 검색 결과 순위를 적용할 필요가 있다. Culture Finder는 시멘틱 웹 검색 에이전트들이 개인화된 문화 정보를 검색할 수 있도록 도움을 준다. Culture Finder는 웹에 존재하는 각 웹 페이지에 대한 메타 데이터를 작성하고, 시멘틱 검색을 이행하며, 사용자 프로파일을 기반으로 삼아 검색 결과에 대한 순위 점수를 계산한다. Culture Finder에는 개인화된 시멘틱 검색을 효율적으로 실행하기 위해 중요한 5가지 기법이 적용되었다.: 사용자의 검색 행위로부터 사용자 프로파일을 생성하기위한 기계 학습기법, 시멘틱 웹 검색 에이전트를 위한 효율적인 시멘틱 검색 기법, 사용자 질의의 효과적인 파악을 위한 질의 분석 기법, 각 사용자에게 적합한 검색 결과를 제공하기 위한 순위 적용 기술, 메타데이터를 생성화기 위한 상위 온톨로지 표현 기법. 본 논문에서는 Culture Finder의 구조를 통해서 시멘틱 개인화 검색에 적용되는 여러 가지 방법을 제안한다.

1. 서 론

현재 시멘틱 웹은 HTML 문서 기반의 웹과 더불어 중요한 웹 환경 패러다임으로 부각되고 있다[1]. 시멘틱 웹은 에이전트, 응용 프로그램, 기업 서버, 웹 커뮤니티들 간의 데이터 공유와 재사용이 가능한 프레임워크를 제공한다[2]. 일반적으로 웹에 분산된 모든 웹 페이지는 구조가 서로 다르다. 시멘틱 웹은 이형적인 구조를 갖는 웹 페이지들의 메타데이터를 바탕으로 시멘틱 검색이 가능하다. 시멘틱 검색은 시멘틱 웹 환경에 맞게 작성된 웹 페이지의 메타 데이터를 바탕으로 실행되며, 메타데이터를 이용하여 웹 페이지의 숨겨진 의미를 추론하여 보다 향상된 검색이 가능하다. 개인화가 적용된 시멘틱 검색은 각 웹 페이지의 메타데이터와 사용자의 성향이 기록된 사용자 프로파일을 사용하여 각 사용자에 맞는 검색 결과를 제공한다. 메타 데이터는 데이터에 관한 구조화된 데이터로, 여기서는 웹 페이지의 특징과 의미를 설명해주는 데이터를 의미한다. 시멘틱 웹에서 중추적인 역할을 하는 온톨로지는 각 웹 페이지에 대한 메타데이터를 구축하기 위한 구조를 제공한다. 따라서 온톨로지를 기반으로 구축된 메타데이터는 사용자의 성향은 학습하기 위한 입력 데이터를 완벽하게 제공한다.

본 논문에서는 문화 정보에 관련된 다양한 웹 페이지로부터 시멘틱 검색을 효과적으로 실행하는 Culture Finder를 설계한다. Culture Finder는 개인화된 시멘틱 검색을 효율적으로 실행하기 위해 중요한 5가지 기법을 적용한다.: 사용자의 검색 행위로부터 사용자 프로파일을 생성하기위한 기계 학습기법, 시멘틱 웹 검색 에이전트를 위한 효율적인 시멘틱 검색 시스템, 사용자 질의의 효과적인 파악을 위한 질의 분석, 각 사용자에게 적합한 검색 결과를 제공하기 위한 순위 적용 기술, 메타데이터를 생성화기 위한 상위 온톨로지 표현.

Culture Finder를 구성하는 시멘틱 개인화 검색 시스템은 키워드나 구문으로 표현된 검색 질의를 시멘틱 검색 엔진에 전송하고, 메타 데이터를 기반으로 개인화가 적용된 검색을 실행하여 그 결과를 정보를 요청한 사용자나 검색 에이전트에 전송한다. 효과적인 시멘틱 개인화 검색을 위해서 본 논문에서는 웹 온톨로지 기술 언어인 OWL을 사용하여 구축된 메타데이터를 기반으로 시멘틱 검색을 하기 위한 방안과 사용자 프로파일 구축 및 검색 결과에 대한 순위 계산 방안을 고려하였다.

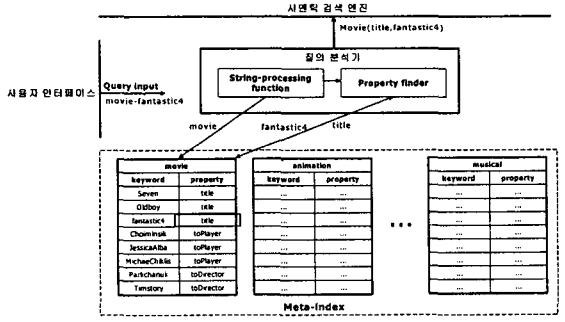
2. 관련연구

RSS[10]과 Atom[11]은 다목적으로 사용하기 위해 적은 용량의 확장된 메타데이터 명세 형식이다. RSS·Atom 형식은 RDF 명세 기준을 따르고 있으며, XML 기반의 응용시스템에 적용된다. 이러한 응용시스템은 웹 사이트의 기본적인 내용을 간결하게 RSS·Atom 형식으로 요약하여 온라인으로 등록하고, 등록된 정보들의 검색 및 전달 또는 회수한다. 시멘틱 웹과 시멘틱 웹 서비스 기술을 통해서 향상된 정보처리를 위한 새로운 가능성과 목표들이 제공되고 있다. 현재, 시멘틱 웹 환경에서 온톨로지 기반의 시멘틱 검색 시스템에 관한 연구들이 진행 중에 있다. Semantic search[3]는 OWL로 작성된 메타데이터 페이지에 존재하는 각각의 객체(Object)들을 검색한다. 따라서 검색 결과는 지식 베이스에 존재하고 있는 객체의 리스트다. Semantic search의 기본 아이디어는 시멘틱 웹상에 존재하는 다양한 페이지들로부터 연관된 데이터만을 획득하는 검색 방식을 채택함으로서, 향상된 메타데이터 검색을 실행하는 것이 목적이다. Swoogle[1]은 키워드를 사용하여 메타 데이터가 저장된 지식 베이스를 검색한다. Swoogle의 검색 결과는 메타데이터 페이지에 존재하는 각각의 객체 뿐 아니라, 주체(subject) 및 속성(property)까지 포함한다. 이러한 종류의 검색은 일반적인 검색 기법을 시멘틱 웹상에 적용한 것으로 볼 수 있다. 그 밖에 시멘틱 웹 기술을 개인 정보 검색 시스템으로 적용한 예로서 Haystack[6]과 Myportal[5]이 있다.

3. 시멘틱 검색 엔진과 질의 분석기

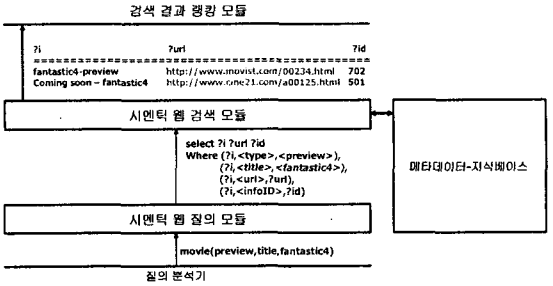
일반적으로 웹 검색 엔진은 키워드와 같이 단순한 질의어를 검색어로 인식하도록 설계된다. 이러한 방식은 각각의 검색 어플리케이션이 질의에 대한 결과를 검색하는데 비효율인 면을 제공한다. 왜냐하면 단순 키워드 매칭은 사용자가 의도하지 않는 검색 결과를 보여주기 때문에 엄청난 양의 정보를 검색한다. 이 경우 수많은 사람이 사용하는 검색 어플리케이션은 검색 엔진의 과부화와 속도 저하현상을 나타내게 된다. 따라서 검색어를 통해 사용자의 검색 의도를 파악하는 것이 중요하다. 검색 질의어의 의미를 파악하기 위해서, 시멘틱 검색 엔진은 메타데이터 기반의 메타색인(Meta-Index)을 사용하는 질의 분석기가 필요하다. 메타데이터는

검색 질의와 검색 결과와의 연결 관계(의미)를 나타내기 때문에, 메타데이터 참조에 의해 구축된 메타색인에는 검색 질의어와 질의어의 의미가 기록된다. 질의 분석기는 두개의 변수(질의 변수, 의미변수)를 사용하여 메타 색인을 인덱싱 하는 과정을 통해서 검색 질의어의 의미를 파악한다. 즉, 질의 분석기는 질의 변수에 저장되는 질의어를 메타색인에 인덱싱 함으로써 질의어가 갖는 의미를 의미 변수를 통해 알아낼 수 있다. 예를 들어 "movie-fantastic4" 라는 질의어가 사용자 또는 검색 에이전트로부터 들어오면 질의 분석기는 movie에 대한 메타색인을 참조함으로써 "movie_Title"이라는 검색 질의어의 의미를 알아낸다. [그림 1]은 질의 분석기가 검색 질의어의 의미를 파악하는 과정을 보여준다.



[그림 1] 검색 질의어 분석 과정

시멘틱 검색 엔진은 시멘틱 웹 관련 기술과 정보 관리 기술을 바탕으로 구현되었다. Culture Finder 적용된 시멘틱 검색 엔진은 메타데이터-지식 베이스(Metadata Knowledge Base), 시멘틱 웹 검색 모듈(Semantic Web Search Module), 시멘틱 웹 질의 모듈(Semantic Web Query Module)로 구성되어 있다. 메타데이터-지식 베이스는 시멘틱 검색에 필요한 메타데이터의 저장소다. 시멘틱 웹 질의 모듈은 질의 분석기로부터 검색 질의어와 질의어의 의미를 전달받아서, 시멘틱 웹 검색 모듈이 처리할 수 있는 형태(RDQL)로 질의문을 재구성한다. RDQL[7]은 그래프 패턴으로 표현이 가능한 트리플 패턴으로 구성된다. 각각의 트리플 패턴은 변수와 RDF 형식의 값(URIs and literals)을 포함한다. 즉, RDF 질의문은 사용자가 요구하는 검색 결과를 바인딩하기 위한 변수들을 선언하고, 트리플 패턴으로 검색 조건을 표현함으로써 시멘틱 웹 검색 모듈이 검색을 실행할 수 있도록 한다. 시멘틱 웹 검색 모듈은 단일화(Unification)을 기반으로 Rete 알고리즘을 사용하여 시멘틱 검색을 실행한다. 단일화는 두 개의 문장이 동일한 표현이 될 수 있도록 변수부호에 항을 대치하는 것을 말한다[8]. 예를 들어서 (∀x) W(x) 문장에서 W(A) 라는 문장을 생성하기 위해서는 변수 x에 문제의 대상 영역에 있는 상수부호 A 가 할당되어야 한다. 즉, W(x) 라는 문장과 W(A) 문장이 동일하게 되기 위해서는 x 가 A로 대치되어야 한다. 이러한 단일화 과정에서 여러 개의 항이 x 에 대치될 수도 있다. 변수를 대응하는 항으로 대체해서 얻어진 표현을 대치문 이라고 하는데, W(A) 는 W(x) 문장에서 얻어진 대치문이 된다. Rete 알고리즘(Forgy, 1982)은 전방향추론 규칙 기반 시스템의 속도를 향상시키기 위한 만들어졌다. Rete 알고리즘은 순환이 없는 방향성 그래프(acyclic directed graph)상에 규칙에 관한 정보를 저장함으로써 속도를 향상시키는 패턴 매치 알고리즘이다. 즉 그래프의 모든 노드 상에서 모든 규칙에 대한 사실(facts)들을 매치시키는 것이 아니고, 변화된 사실들에 대해서만 매치 시킨다. 따라서 각 노드에서 변화가 없었던 정적인 데이터는 무시되기 때문에 사실들의 매치 속도는 크게 향상된다. Rete 알고리즘이 적용된 시멘틱 웹 검색 모듈은 질의문의 트리플 패턴들을 하나씩 분리하여 루트 노드를 제외한 노드들에 표현한다. 즉, 루트 노드로부터 최하위 노드까지의 경로를 통해서 질의문의 모든 트리플 패턴들이 표현되며 루트 노드로부터 현 위치의 노드까지 각 경로에 저장된 질의 패턴을 만족하는 메타데이터에 대한 정보를 각각의 노드에 저장한다. 이 정보는 패턴내의 변수에 대치되는 모든 값(Values)들을 나타낸다. 이것은 시간 중복성(Temporal Redundancy)을 줄여주는 장점이 있다. 따라서 루트 노드로부터 경로가 진행될수록 보다 적은 메타데이터가 매치된다. [그림 2]는 메타데이터를 기반으로 시멘틱 검색을 실행하는 과정을 보여준다.



[그림 2] 시멘틱 검색 실행 과정

4. 사용자 프로파일 학습 및 검색 결과 랭킹

사용자 행위 모니터 모듈은 시멘틱 검색 시스템을 사용하는 사용자 개개인의 검색 행위(무엇을 검색하고, 어떤 검색 결과에 관심을 가졌으며, 각 검색 결과가 가지는 특징)에 대한 데이터를 기록한다. 일반적으로 웹 포털 서비스를 이용하는 모든 사용자의 행위는 포털 서비스를 제공하는 웹 서버의 로그에 기록된다. 그러나 로그 정보는 오직 사용자가 관심(마우스 클릭에 의한 선택이나 방문)을 가지는 페이지의 주소에 대한 정보만을 포함하고 있기 때문에, 사용자의 검색 행위를 학습하기에는 충분하지 못하다. 이러한 문제점을 해결하기 위해서 사용자 행위 모니터 모듈은 웹 서버의 로그 정보를 바탕으로 각 사용자의 모니터 테이블을 만든다.

사용자 모니터 테이블을 구성하기 위한 필수 조건은 검색 결과에 대한 마우스 클릭 정보와 메타데이터 ID이다. 메타데이터 ID는 검색된 각각의 페이지와 이러한 페이지들의 메타데이터를 연결해 주는 매개체다. 사용자 모니터 테이블은 사용자 성향 학습 엔진이 사용자의 정보 검색 성향을 학습하기 위한 상세한 데이터(웹 페이지에 대한 여러 가지 속성 및 속성 값)를 제공한다. 검색된 페이지의 상세한 데이터는 온톨로지를 기반으로 생성된 메타데이터에 대응되기 때문에, 사용자의 정보 검색 성향을 학습하기 위한 상세한 데이터는 각각의 페이지에 대한 메타데이터로부터 메타데이터 ID를 통해서 생성된다. 다음 [표 1]은 Culture Finder에 사용된 사용자 모니터 테이블을 생성하기 위해 정의한 알고리즘이다.

```
• Define learning field based on ontology
• for(fetchloopsizeA=0; fetchloopsizeA<size of user table;
      fetchloopsizeA++)
   • User ID fetch in user table(fetch(fetchloopsizeA))
   • Make user's monitor table to name user ID
   • for(fetchloopsizeB=0; fetchloopsizeB<size of log table;
         fetchloopsizeB++)
      • Category, subcategory, and information ID fetch in log
        table(fetch(fetchloopsizeB))
      • Detailed data load in meta-data repository
        (load(category, subcategory, information ID))
      • Record detailed data in learning field
```

[표 1] 사용자 모니터 테이블 생성 알고리즘

사용자 모니터 테이블은 사용자가 관심을 갖는 검색 결과들을 사용자의 직접적인 개입 없이 마우스 클릭 유무로서 표현한다. 따라서 시멘틱 개인화 검색을 실행하는 Culture Finder는 각 사용자의 정보 성향에 맞는 검색 결과를 우선적으로 보여주기 위해, 사용자 모니터 테이블을 기반으로 사용자 프로파일을 생성한다.

사용자 프로파일을 구축하는 방식에는 지식 기반 프로파일 구축 방법과 행위 기반 프로파일 구축 방법이 있다. 지식 기반 프로파일 구축 방법은 엔지니어가 미리 사용자들의 프로파일 모델들을 구현

하고, 설문 조사와 인터뷰를 통해서 실제 서비스를 이용하는 사용자와 가장 가까운 모델을 그 사용자의 프로파일로 선택한다. 행위 기반 프로파일 구축 방법은 보통 기계 학습(Machine Learning)기법을 사용하여 사용자 행위의 패턴을 학습하고, 이러한 행위 패턴을 기반으로 프로파일을 구축한다. 본 논문에서 제안하는 Culture Finder의 사용자 프로파일 생성기는 행위 기반 프로파일 구축 기법을 기반으로 한다. 사용자 모니터 테이블에 저장된 상세 데이터는 검색된 웹 페이지의 여러 특성을 나타내는 속성과 속성 값이 저장된다. 따라서 사용자 프로파일 생성기는 각 사용자의 모니터 테이블과 사용자 성향 학습 알고리즘을 기반으로, 페이지 정보를 구성하는 각 속성마다 사용자의 기호를 나타내는 값들이 기록된 사용자 프로파일을 생성한다.

사용자 프로파일은 특정 주기(하루 또는 일주일에 한번)마다 새로 생성되는데, 이 때 사용자의 피드백이 큰 변수로 작용한다. 즉, 순위가 적용된 검색 결과 페이지(추천 검색 페이지)에 사용자가 관심을 갖지 않을 경우, 그 페이지를 구성하는 속성 값과 일치하는 프로파일의 속성 값 가중치를 낮추어야 하는 반면, 순위가 적용되지 않은 검색 결과 페이지(비 추천 검색 페이지)에 사용자가 관심을 갖는 경우, 그 페이지를 구성하는 속성 값과 일치하는 프로파일의 속성 값 가중치를 높여 주어야 한다. 사용자 피드백 정보 역시 순위가 적용된 결과에 대한 사용자의 마우스 클릭 유무로 제공받으며, 모든 피드백 정보는 로그 테이블에 저장된다. 따라서 사용자 프로파일 생성 모듈은 모니터 테이블의 상세 데이터를 기반으로 사용자 관심 속성 값들의 가중치를 계산하고, 사용자 피드백 정보를 바탕으로 관심 속성 값들의 가중치를 자동으로 조절하여 최종적으로 사용자의 성향이 반영된 사용자 프로파일을 생성한다.

Culture Finder에서 사용되는 사용자 프로파일에는 각각의 문화 정보(영화, 애니메이션, 음악 등등)의 세부 분류 정보(영화 리뷰, 인터뷰, 전문가 영화평 등등)에 대한 각 사용자의 관심 속성 값의 가중치가 저장됨으로서, 사용자의 문화 정보 성향을 파악할 수 있는 단서를 제공한다. Culture Finder의 사용자 프로파일 생성 모듈이 관심 속성 값의 가중치를 계산할 수 있도록, 다음과 같은 확률 기반의 가중치 계산식(WNP funct ion)을 정의하였다.

Feedback Weight = [positive-weight, negative-weight]
WNP ratio = NP ratio + ((positive ratio*positive-weight) + (negative ratio * negative-weight))

NP ratio = Frequency ratio * positive ratio
Positive ratio = positive Count(attribute value) / tot_Count(attribute value)
Frequency ratio = positive Count(attribute value) / ∑positive Count(attribute)

Culture Finder는 검색된 정보의 분류에 따라서 속성 값들의 가중치가 정해지며, 일정 기준(Weight_Ordering threshold) 이상의 가중치를 갖는 속성 값을 선택하여 사용자 프로파일을 구축하게 된다. 사용자 프로파일은 각 사용자의 ID명과 일치하는 프로파일 테이블들로 구축된다. 본 연구에서 사용자 프로파일을 구축하는 방법은 다음과 같이 정리할 수 있다. 모니터 테이블을 바탕으로 각 속성별로 속성 값들의 가중치를 구하고, 가중치가 높은 순으로 정렬한 후 일정 가중치를 넘는 속성 값들만 프로파일 테이블의 해당 속성 필드에 삽입한다.

검색 결과 랭킹 모듈은 사용자 프로파일을 참조하여 검색 결과에 추천 점수(Recommend Confidence)를 할당한다. 일정한 기준 이상의 추천 점수(Recommend Confidence)를 갖는 검색 결과는 추천 점수가 큰 순으로 순위가 적용되어 시멘틱 개인화 검색 결과로 제공된다. 즉, 순위가 적용된 정보는 사용자의 정보 성향이 반영된 시멘틱 개인화 검색 결과이며, 순위가 적용되지 않은 정보는 일반적인 시멘틱 검색 결과다. 추천 점수는 랭커(Ranker) 알고리즘에 의해 계산되며, [표 2]는 Culture Finder의 검색 결과 랭킹 모듈이 검색 결과들의 추천 점수를 계산하여 순위를 적용하기 위한 알고리즘을 정의한 것이다.

```
• for(loopsize=0; loopsize<number of search results; loopsize++)
  • information ID fetch from search results(fetch(loopsize))
  • for(fetchloopsizeC=0; fetchloopsizeC<number of attribute;
      fetchloopsizeC++)
    • A value of attribute fetch in meta-data DB
      (fetch(fetchloopsizeC, information ID, attribute))
    • Interest values fetch in user profile
      (fetch(attribute, userID))
      • if(first interest value == value)
          confidence point = 1st_point(1.0)
      • else if(second interest value == value)
          confidence point = 2st_point(0.8)
      • else if(third interest value == value)
          confidence point = 3st_point(0.5)
      • else confidence point = zero
    • rank_point = rank_point + confidence point
  • if(rank_point >= threshold)
      assign recommend confidence (rank_point) to search
      result
  • else
      assign recommend confidence(zero) to search result
• Sort search result according to recommend confidence
```

[표 2] 검색 결과 순위 적용 알고리즘

5. 결론 및 향후 연구

본 논문에서는, 문화 정보에 대한 시멘틱 개인화 검색을 실행하는 Culture Finder의 구조를 제안하고, 이러한 구조를 통해서 시멘틱 개인화 검색에 적용되는 여러 가지 기술에 대해 설명하였다. Culture Finder를 설계하기 위해서, 본 논문에서는 웹 온톨로지 기술 언어인 OWL을 기반으로 효율적으로 메타데이터를 구축하기 위한 방안, 질의 분석 및 시멘틱 검색 방법과 사용자 프로파일 구축 및 검색 결과에 대한 순위 계산 방안을 고려하였다. Culture Finder는 시멘틱 웹 검색 에이전트가 문화(문화 뉴스, 공연, 전시)에 관련된 다양한 정보들에 대해서 개인화된 검색을 하는데 도움을 준다. 본 논문에서 제안된 구조는 웹 콘텐츠 및 서비스를 일괄적으로 자동(또는 반자동)통합하여, 웹상에 존재하는 자원을 정확하게 사용하는 시멘틱 웹, 웹 서비스 및 다중 에이전트 기술에 적용될 수 있다. 본 논문에서 제안된 시멘틱 개인화 검색 시스템은 현재 계속 진행 중인 연구로서, 향후에는 온톨로지 기반의 사용자의 정보 성향 학습과 사용자 프로파일을 기반의 검색 결과 랭킹 적용 기법에 대해서 보다 자세한 연구가 진행될 것이다.

참고문헌

[1] Li Ding, Tim Finin, "Swoogle: A Search and Metadata Engine for the semantic web", CIKM'04, November 8-13, 2004, Washington DC, USA
[2] Semantic web, http://www.w3.org/2001/sw/
[3] Guha, R. V,. McCool, R. and Miller, E. "Semantic search", Proceedings of the twelfth international conference on World Wide Web(WWW203), ACM Press, 2003
[4] MSRBot,http://www.research.microsoft.com/research/sv/msrbot/#webcrawler
[5] Haibo Yu, Tsunenori Mine, Makoto Amamiya, "An Architecture for Personal Semantic Web Information Retrieval System", WWW 2005, May 10~14, 2005, Chiba, Japan
[6] D. Quan and D.R.karger, "How to make a semantic web browser", WWW 2004
[7] RDQL - A Query Language for RDF, http://www.w3.org/Submission/2004 /SUBM-RDQL-20040109/, January 2004
[8] George F Luger, "ARTIFICIAL INTELLIGENCE", Addison-Wesley, pp 67~68,1994
[9] The RETE Algorithm, http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/
[10] RDF Site Summary (RSS)1.0, http://web.resource.org/rss/1.0/
[11] Atom, http://www.mnot.net/drafts/draft-nottingham-atom-format-02.html
[12] Stuart. E. Middleton, Nigel. R. Shadbolt, David C de Roure, "Ontological User Profiling in Recommender System", ACM Transaction on Information System, Vol. 22, No. 1, January 2004, pp 54~88.
[13] OntoMat-Annotizer, http://annotation.semanticweb.org/ontomat/index.html
[14] Welcome to protege, http://protege.stanford.edu/index.html