웹 상에서의 특정 장르 문서 발견

주원균* **, 맹성현**

*연구개발정보센터(KORDIC), **충남대학교 컴퓨터과학과

Discovery of Genre Information on the Web

Won-Kyun Joo* **, Sung Hyon Myaeng**

*KORDIC (KOrea Research and Development Information Center)
**Dept. of Computer Science, Chungnam National University

요 약

정보공유를 목적으로 제안된 웹의 활성화와 함께 유용한 정보들이 웹상에 기하급수적으로 등장함에 따라 정보공간의 확장으로 인한 검색 신뢰도의 저하 문제에 직면하게 되었다. 본 연구에서는 대용량 웹 환경하에서 사용자의 정보발견을 돕기 위해 텍스트이외의 새로운 요소들을 사용하여 특정장르문서를 발견하는 개념을 도입하였다. 먼저 사용자가 발견하고자 하는 장르의 모습을 텍스트, URL정보, 링크 정보, 문서구조 정보 등의 장르 식별요소 값을 이용해 표현한 후, 후보 문서들의 장르관련도를 측정함으로써 특정장르 문서를 검색한다. 각 장르식별요소값은 나름대로의 방법에 의해 계산되는데 0~1사이의 값을 가지며, 종합적인 장르관련도는 각 장르식별요소값의 증거통합 방법에 의해 구한다.

본 논문에서는 각 장르식별요소들의 역할과 장르식별요소가 장르발견에 미치는 영향을 알아보며, 최종적으로 특정 장르 문서발견에 있어서의 검색 신뢰도 향상을 보이기 위해 실험모델을 설계/구현하였다. 본 실험은 웹 문서를 대상으로 하는데, 아직까지 URL, 링크 정보를 모두 갖춘 테스트컬렉션이 없기 때문에 실험을 위해 일반적인 웹 문서로 직접 구성한 컬렉션을 사용하였다. 발견하고자 하는 장르는 "컴퓨터 분야의 컨퍼런스 홈페이지"로 정하였으며 30개의 컴퓨터 분야를 선정하였다. 비교대상으로는 일반 웹 검색 엔진인 알타비스타와 메타검색 엔진인 메타크롤러를 선택하였고, 각 질의에 대해 상위 30개의 결과를 대상으로 정확도를 평가하였다. 결과로서 각 장르식별요소들은 모두 검색 신뢰도의 향상에 기여를 하며, 제안하는 방법은 알타비스타와 메타크롤러에 비해 각각 평균적으로 67.34%, 71.78%의 검색 신뢰도 향상을 보임을 입증하였다.

1. 소개

정보공유를 목적으로 제안된 웹의 활성화와 함께 접하기 어려웠던 많은 유용한 정보들이 웹상에 등장함에 따라 웹정보검색은 새로운 국면을 맞이하게 되었다. 웹 검색의 초창기에는 문서의 규모가 비교적 작았으며, 사용자도 대부분 만족할 만한 결과를 얻을 수 있었다. 그러나 한 통계 자료[14]에 따르면, 1994년에 최초의 웹 검색 엔진인 World Wide Web Worm이 11만개의 웹 페이지들에 대한 정보를 제공한데 비하여, 현 상황 (1999년 5월 1일)에서는 200만 (WebCrawler)에서 15,000만개 (AltaVista)의 서비스 가능한 웹 문서가 존재한다. 이 수치는 웹의 규모를 대변해주는데, 이처럼 웹의 대규모화로 인해 검색의 여러 면에서 많은 어려움에 직면하고 있으며, 특히 검색 신뢰도의 문제가 심각하다[15]. 한 예로 사용자가 하이퍼텍스 컨퍼런스(hypertext conference) 관련 홈페이지를 발견하기 위해 알타비스타를 이용하여 "hypertext conference"라는 질의를 통해 검색 했을 때 결과로서 약 100만 건의 문서가 된다.

대용량 정보환경에서의 정보발견을 돕기 위해서 자원발견(Resource Discovery), 정보발견(Information Discovery), 데이터마이닝(Data Mining)과 같은 개념들이 등장했다. 정보발견에 대한 정의를 통해 정보발견의 범주와 필요성을 파악할 수 있는데, Yeong는 "발견, 검색, 전달[19]"이라는 용어를 빌어 설명하였고, Bowman외는 "정보 인터페이스, 분산, 수집[2]"이라는 말로, Deutsch는 두 개념을 조합하여 "클래스 발견, 인스턴스의 위치, 인스턴스에 대한 접근, 정보 관리[5]"라는 말로 표현하고 있다. Iannella는 종합적인 관점에서 자원, 발견, 사용자, 서비스 공급자의 네 주체가 명시되고 이들 사이의 특성을 정의함으로써 정보발견에 대한 정의를 내렸다[8]. 결국 정보발견이라는 분야를 사용자 관점에서 해석한다면, 분산 환경에서 서비스의 주체인 사용자의 요구를 최대한으로 반영하기 위해 다양한 요소들을 활용하고, 정확한 해석을 통해 사용자에게 최적의 정보를 제공해 주기 위한 방법이다.

지금까지 정보발견에 관련된 다양한 연구들이 진행되어 왔는데, 본 논문에서는 이용하는 정보의 종류와 방법에 따라 링크 기반, 구조 기반, 소프트봇 기반으로 나눈다.

- 링크 기반의 연구는 링크가 관련 있는 두 문서를 연결한다는 링크 특성[4,10,20]을 이용하여 링크의 영향력을 정보 검색 결과에 반영시킴으로써 검색 신뢰도의 향상[9,23] 또는 특정 주제에 대한 발견[11,16]을 도와준다.
- 구조 기반의 연구는 HTML 혹은 링크에 의한 구조를 사용한다. HTML기반의 연구는 HTML이 중요 태그에 의한 부분구조를 가지고 있다는 특성을 활용하는 방법[3]이며, 링크에 의한 구조는 링크의 전승(predecessor), 계승(successor) 관계를 반영함으로써 특정 주제에 대한 발견을 효율을 높일 수 있는 방법[7]이다. 또한 유사한 필드 구조를 지닌 웹 검색 결과의 융합에 의해 검색 신뢰도를 향상시키는 기법[17]도 연구되었다.
- 소프트봇[12] 기반의 연구는 인공지능을 갖춘 에이전트를 기반으로 외부의 사용 가능한 모든 자원들을 이용하여 검색/추론/지식 발견의 과정을 거쳐 사용자의 정보 발견 요구를 만족시켜 주는 것을 목표로 하는데, 사용자의 홈페이지의 발견을 도와주는 Ahoy![9], 웹 상에 존재하는 과학관련 논문의 발견을 도와주는 WEBFIND[1], 인터넷 상의 FAQ의 특정 항목에 대한 발견을 위한 FAQ Finder[13]가 있다.

하이퍼텍스트 문서를 가진 웹 환경에서는 기존의 텍스트 환경에 비해 URL, 문서간 링크, 문서 구조 등의 새로운 요소들이 출현하였고 이 요소를 잘 활용함으로써 정보발견의 신뢰도를 높일 수 있다. 기존의 연구들은 모두 정보발견에 새로운 요소를 반영하려는 시도들이다. 본 논문은 웹 문서의 새로운 요소들을 고려하여 사용자 측면에 맞춘 장르라는 개념을 정의하고, 이 특정 장르 웹문서를 발견하는 것을 목표로 한다. 검색 신뢰도 측면에서 효과가 이미 입증된 링크의 영향[18]과 문서 구조 이외의 URL 패턴에 따른 영향력을 추가하여, 이들 증거들에 대한 종합적인 해석을 통해 최종 장르 발견에 도달한다.

본 논문은 다음과 같은 구조를 가지는데, 2장에서는 장르에 대한 정의와 장르를 식별하기 위한 요소들에 대해 언급하며, 3장에서는 장르 식별을 위한 장르 식별 요소의 추출 및 장르 식별 알고리즘에 대해, 4장에서는 실험을 통해 각 요소들의 영향과 정보발견 신뢰도의 향상에 관해 보이며, 5장에서 결론을 맺는다.

2. 장르 발견

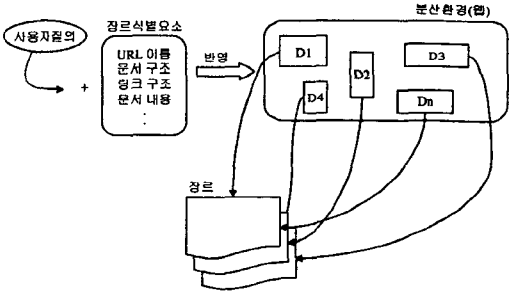
장르발견은 단순 키워드에 기반한 기존의 정보검색방법을 탈피하여 웹과 같은 분산환경의 다양한 특성들을 정보 검색에 활용하여 사용자의 요구를 보다 능동적으로 반영함으로써 정보발견의 신뢰도를 높일 수 있는 방법이다. 예를 들면 특정내용이나 구조를 지닌 컨퍼런스 개최 홈페이지의 발견, 특정 분야에 관련된 사용자 홈페이지의 발견과 같은 것들이다.

2.1. 장르의 정의

장르는 특정 종류 또는 유형의 문서집합을 일컫는 용어로서 해당 장르에 속하는 문서의 종류/유형을 구분하기 위해 장르 특성을 사용한다. 장르 특성으로는 웹의 URL패턴, 웹 문서의 구조, 웹 상의 링크 구조 등이 될 수 있다. 사용자마다 발견하고자 하는 정보가 다르므로 장르는 사용자마다 다르게 정의 될 수 있는데, 장르는 일종의 사용자 관점에 맞추어진 검색 결과 도메인이라 할 수 있다. 그러나 일반적인 도메인이 정보원의 텍스트 맥락에서 특정 내용 에 의한 구분이라면, 장르는 그 문서의 용도 혹은 유형 분류에 초점을 맞춘 구분이고, 텍스트뿐만 아니라 문서의 다른 요소를 추가하여 식별한다는 점에서 차이를 보이고 있다. 따라서 장르별 질의는 웹의 새로운 요소를 추가함으로써 사용자 질의를 확장한 형태를 보이고 있다.

그림1은 위에서 설명한 장르를 사용한 웹 문서 검색의 개념을 도식으로 보인다. 여기서 웹의 다양한 특성은 장르특성으로 표현되어 있고, 분산 환경의 문서 혹은 도메인은 D<1-n>으로 표현되어 있다.

장르는 장르식별요소를 정보 발견에 반영함으로써 의미를 가지는데, 장르식별요소는 종전의 텍스트로 명시된 질의의 부족함을 보완/확장하기 위한 방법으로서 정보발견 시 웹 환경의 특성을 충분히 활용하기 위한 수단이다. 따라서 사용자가 특정 장르를 발견하고자 할 경우 이에 상응하는 장르식별요소를 명시함으로써 웹 환경의 특성을 충분히 수용하여 원하는 장르를 발견해 낼 수 있다.



[그림1] 장르에 대한 정의

2.2. 장르식별 요소

장르 발견의 관건은 얼마나 유용한 장르식별요소를 사용하는가에 달려있다. 종전의 텍스트 문서에서는 텍스트 이외의 정보들이 많지 않았지만 분산 하이퍼 텍스트 환경으로 전환되면서 정보발견에 도움이 될 수 있는 유용한 정보들이 많이 생겼다. 본 논문은 웹상에서 장르식별요소로 사용될 수 있는 것에 URL패턴, 문서의 구조, 링크 패턴, 문서 내용을 포함한다.

■ URL 패턴

URL패턴은 웹 문서의 위치를 유일하게 결정지어 주는 이름으로서 웹 문서의 저작자들이 문서의 특성과 관련된 패턴을 문서URL 이름으로 부여한다는 가정을 전제로 한다. 예를 들어 사용자가 98년도에 열린 하이퍼텍스트 컨퍼런스의 페이지의 URL을 결정할 때 다음과 같은 패턴을 사용할 가능성이 높다.

http://ht98.chungnam.ac.kr/
http://hypertext98.chungnam.ac.kr/
http://irsun.chungnam.ac.kr/ht98/
http://irsun.chungnam.ac.kr/ht98/homepage.htm(l)
http://irsun.chungnam.ac.kr/ht98/index.htm(l)
http://irsun.chungnam.ac.kr/hypertext/
http://irsun.chungnam.ac.kr/hypertext98/
http://irsun.chungnam.ac.kr/hconf/
http://irsun.chungnam.ac.kr/conf/hypertext98.html

이러한 특성은 사용자 질의와 URL 패턴과의 관계를 파악함으로써 장르 발견에 도움을 줄 수 있다는 것을 의미한다.

■ 문서구조

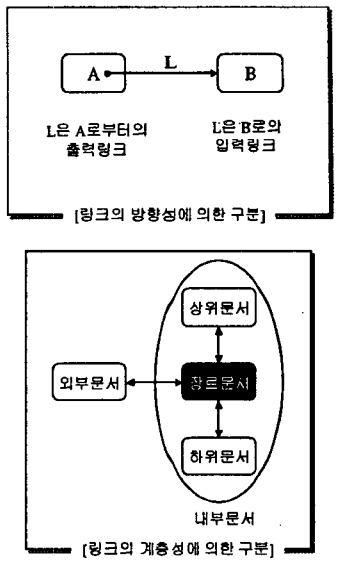
문서 구조에 대한 정의와 여러 태그들의 작용에 의해서 생성된 SGML과 같은 종류의 문서는 구조를 지닌 문서에 포함시킬 수 있다. 본 논문에서 대상으로 하는 HTML문서(웹 문서)는 구조적인 문서 라기 보다는 표현을 위한 언어지만, 부분적으로 구조를 지니고 있어서 그 문서 구조를 장르발견에 활용할 수 있다. HTML문서의 부분 구조는 <H1>~<H6>, <CENTER>, <P>,
 등의 태그와 사용자의 의미 표현 능력에 의해 나타난다. 구조 정보의 형태 및 내용은 문서에 따라 차이가 있겠지만 대상문서를 컨퍼런스 관련 문서라 가정한다면 타이틀, 로고, 개최 날짜, 개최 장소, 프로그램, 프로그램 의장 등에 해당하는 내용이 구조화되어 표현 될 수 있다.

저작자들은 타이틀의 표현을 위해 <TITLE>혹은 <H1-H3>에 의한 문단을 문서의 앞쪽에 위치 시키는 경향이 있으며, 로고는 <IMG>태그를 사용하여 문서의 앞쪽에 위치시키는 것이 일반적이다. 마찬가지로 개최 날짜의 경우는 <H1> ~ <H3>, CENTER등의 태그를 사용하여 문서의 앞쪽에 보이는 것이 일반적이다. 그림2는 이를 설명하고 있다.

■ 링크패턴

링크는 웹 상의 가장 중요한 요소중의 하나로써 문서들을 복잡한 그래프 형태로 연결한다. 링크의 영향력의 반영은 링크를 보는 관점에 따라 다른데, 본 논문의 관점에서 해석한다면 그림2와 같은 링크 패턴을 찾을 수 있을 것이다.

그림2의 위쪽은 링크를 방향성에 의해 구분한 것이며 아래쪽은 특정 장르문서를 중심으로 링크의 계층성에 의해 외부/내부 문서, 상위/하위 문서로 구분한 것이다. 링크의 방향성은 한 문서를 중심으로 다른 문서로의 출력인지 입력인지의 구분에 따른다. 링크의 계층성은 링크의 URL 이름을 사용하여 구분하는데, 우선 URL이름을 호스트, 패스, 파일이름으로 분리한다. 분리된 URL에 대해서는 장르문서라고 추측되는 문서의 URL과 링크로 연결된 문서들의 URL을 비교하는데, 호스트 부분이 다른지의 여부에 따라 외부 문서와 내부 문서로 나눈다. 호스트 부분이 같은 내부 문서에 대해서는 패스 부분이 장르 문서 패스의 포함여부에 따라 완전 경로를 포함하면 하위 문서, 일부를 포함하면 상위 문서로 나눈다.



[그림 2] 링크의 방향성과 계층성의 의한 구분

링크 패턴에 따른 문서 참조 빈도 수를 장르발견에 사용하기 위해서 링크의 영향력은 외부문서로의 출력링크, 외부문서로부터의 입력 링크, 상위문서로의 출력링크, 상위문서로부터의 입력 링크, 하위문서로의 출력링크, 하위문서로부터의 입력링크의 6가지로 구분할 수 있다.

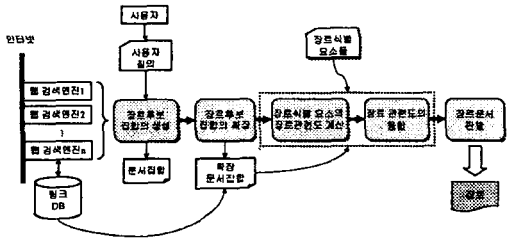
■ 문서내용

문서 내용은 텍스트 기반의 정보검색[22]에서 사용하는 요소로 문서와 질의어와의 관련도를 측정하기 위해서, 사용자가 입력한 질의어에 대한 대상 문서의 단어 빈도(Term Frequency)와 역문서빈도(Inverted Document Frequency)를 주로 이용한다.

본 논문에서는 크게 4종류의 장르식별 요소들과 그것들에 대한 세부 내용을 사용하여 장르별 문서 검색 과정을 설명하는데, 이러한 요소들과 세부항목은 응용 분야 혹은 장르에 따라 재정의 혹은 변경되어야 한다. 만약 사용자가 '사용자 홈페이지'라는 특정 장르의 발견하고자 한다면 '사용자 이름'이라는 세부항목을 문서 구조 요소에 자유롭게 추가할 수 있다.

3. 장르 발견 알고리즘

일반적인 장르발견 알고리즘은 다음과 같은 단계를 가진다.



[그림3] 장르 발견 과정

A. 사용자들로부터 질의를 입력 받아 분산환경의 웹 검색 엔진들의 상위 결과를 취합하여 장르 발견의 출발점이 되는 장르 문서 후보 집합을 결정한다.
B. 장르 문서 후보 집합에 링크로 연결 된 문서를 포함하여 확장 문서 집합을 생성한다.
C. 확장 문서 집합의 모든 문서들에 대해 사용자가 선택한 장르에 적합한 식별 템플릿을 사용하여 장르 식별 요소를 추출하고 각 요소들의 장르 관련도를 융합하여 각 문서의 장르 관련도를 구한다.
D. 상위 몇 개 혹은 특정 임계값 이상의 값을 가지는 문서들로 장르를 구성하여 사용자에게 제시한다.

3.1. 장르 문서 후보 집합의 결정

장르발견의 첫 단계로 사용자가 요청한 장르의 후보 문서를 수집하는 과정이다. 사용자 질의를 검색어로 사용하여 여러 웹 검색 엔진들에 요청하여 각 엔진별로 상위 k개의 문서를 취한다[1]. 실험에 의하면 k는 30일 때 가장 만족할 만한 결과를 보인다(4장 참조). 각 검색 엔진들의 결과 문서 수집을 위해서 자바기반의 멀티 쓰레드로 설계/구현 된 문서수집 로봇[21]을 사용하였다.

3.2. 장르 문서 후보 집합의 확장

웹 검색 엔진들에 대한 검색의 결과로서 생성된 장르 문서 후보 집합을 확장하는데 이것은 웹 문서의 특징을 활용하여 재현율을 향상시키기 위함이다. 링크는 서로 관련 있는 문서를 연결한다는 특성[4,10,20]을 지니고 있는데, 링크에 의해 후보 집합의 문서와 연결된 새 문서들도 결과로 포함시킴으로써 장르 문서 발견 확률을 높일 수 있으며, 검색 엔진들에 등록 되지 않은 문서를 제시할 수 있다. 이것은 질의 확장과 같은 역할을 한다.

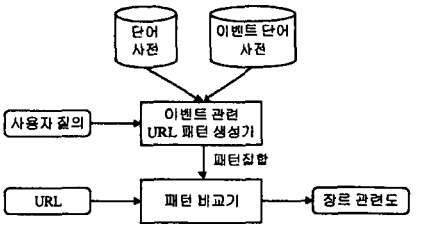
이 과정에서는 부수적으로 dangling link를 제거할 수 있다. 웹의 방대함으로 인해 웹 검색 엔진은 비교적 긴 문서 갱신 주기를 가진다. 시스템에 따라 다르지만 국내의 경우 대략 한달 정도의 갱신주기를 가진다. 그러나 사용자의 문서 갱신 주기는 비교적 빈번하므로 웹 검색 결과에서 많은 dangling link가 발생한다. 일단 확장을 위해서는 장르 후보 문서들을 수집하는 과정이 필요한데 이 과정에서 존재하지 않는 문서를 제거함으로써 사용자의 결과 사용 시간을 줄이며 보다 신뢰성 있는 정보를 제공할 수 있다.

반면에 링크 확장으로 인해 후보 문서집합의 크기가 커지는데, 이것은 장르 식별 시간과 오류를 가중시킨다. 따라서 불필요한 링크의 확장을 배제시킬 수 있는 방법과 어느 정도의 확장을 취해야 하는가에 대한 연구가 필요하다.

3.3. 장르식별 요소로부터의 관련도 추출

3.3.1. URL패턴의 장르 관련값

URL패턴으로부터 장르 관련값을 구하기 위해 질의와 URL과의 패턴매칭 방법을 사용하는데 전체적인 모습은 그림4에 보인다.



[그림4] URL패턴의 추출 및 장르 관련값 측정

먼저 사용자 질의를 바탕으로 하여 가능한 URL패턴 집합을 생성해 내고 검색대상 문서의 URL과 패턴 집합들과의 관련도를 계산하는데, 이 값이 URL패턴의 장르 관련값이다.

그림4에서 패턴 생성기는 사용자 질의를 사용하여 가능한 이벤트 URL패턴 집합을 생성해 낸다. 패턴 비교기는 장르후보집합의 URL들에 대해 패턴 집합과의 비교를 통해 URL에 의한 장르 관련도를 구한다. 패턴생성시에는 명사 및 접두사 사전을 사용하는데, 사용자 질의의 각 단어들을 기본단위로 잘라내고 다시 이를 조합함으로써 패턴들이 생성된다. 이때 단어 사전은 약 4만개의 영어 단어에 대한 정보를 가지고 있는 사전이며, 이벤트 단어 사전은 이벤트 관련 단어들(DL, SIGIR 등)에 대한 정보를 가지고 있는 사전으로 접두사 및 이벤트 단어 사전은 지속적으로 확장되고 있다. 생성되는 패턴유형은 다음과 같다.

① 질의 각 단어의 전체 혹은 부분단어(prefix등)
② 질의 단어들의 전체/부분/첫이름(Initial)의 조합
③ 위의 것들과 숫자의 조합형태

다음은 이해를 돕기 위해 사용자 질의로 'conference hypertext'가 주어 졌을 때 패턴 유형의 각각에 대한 예이다.

① conference, hypertext, conf, hyper
② hypertextconference, hyperconf, hconf, ht, hc
③ hyper98, ht98, hc98, etc.

패턴 생성기에서 생성된 패턴들을 주어진 URL에 적용시켜, 수식1에 따라 관련도를 부여한다.

[1] 본 연구에서는 검색 신뢰도 향상을 위해 분산검색 기법을 사용하고 있으나 하나의 특정 검색엔진을 사용하는 경우에도 동일 알고리즘을 사용할 수 있다.

$$E_u = \begin{cases} 0 : URL\text{이 패턴을 포함하지 않을 경우} \\ 1 : URL\text{이 패턴을 포함 할 경우} \end{cases} \quad (1)$$

패턴의 포함여부를 결정하기 위해서 다음의 규칙을 적용시킨다.

① 패턴집합은 URL의 호스트 이름과 경로상의 단어들에 대해 적용시키며, 각 단어들의 구분을 위한 구분자로서 '.', '/' 을 사용한다.
② 패턴은 각 단어의 시작 부분에 위치한다.

3.3.2. 문서 구조의 장르 관련값

추출하고자 하는 구조 정보로는 장르 문서 제목, 로고, 장소 및 날짜가 있다. 이러한 구조 정보의 존재 여부를 파악하기 위해서 본 논문에서는 HTML의 태그 특성을 이용한 패턴 매칭 방법을 사용한다. 패턴 매칭의 범위를 줄이기 위해서 다음의 HTML태그 특성을 이용하여 중요 문장을 뽑아낸다.

① 중요문장에는 <H1>~<H3> 태그 혹은 <CENTER>와의 조합에 의한 태그를 많이 사용한다.
② <LI>, <DD>등의 항목 나열자를 많이 사용한다.

추출된 중요문장에 대해서는 다음 알고리즘을 적용하여 문서 구조정보의 포함여부를 확인한다.

A. 중요문장의 각각에 대해서 초기값으로 0을 부여한다.
B. 중요문장에 대해서 표2의 적용 기준들에 의해 각각의 구조 정보 가치를 부여한다. 이때 적합 기준에는 +의 가치를 비적합에는 – 의 가치를 부여한다.
C. 중요문장의 각각에 대해서 구조 정보의 가치가 임계값(p) 이하인 것은 제거한다.
D. 최종적으로 각 구조정보의 가치가 큰 가장 중요문장을 그에 해당하는 구조 정보로 삼는다.

[표 2] 적용기준

구조 정보	적용 기준	
	적 합	비 적 합
장르 제목	-질의 문자열의 포함 -서수 형태의 포함 -전치사 on, of, 's의 포함	- 질의 문자열의 비포함
로고	-이미지 URL의 질의 관련성 -ALT문자열의 질의포함	
장소	-지명의 포함 여부 -','에 의한 문자열의 나열 정도	
날짜	-연도 정보의 포함 -'-'에 의한 숫자열의 포함	

구조 정보에 의한 장르 관련도는 수식2에 의해 반영한다.

$$E_s = \frac{\sum_{i=1}^{n} S_i}{n} \quad (n : \text{포함 할 구조 정보 가치의 수}) \quad (2)$$
$$S = \{\text{이벤트 제목, 이벤트 로고, 개최 날짜, 개최 장소}\}$$

이때 S_i는 해당 문서 구조정보의 출현여부, i는 집합 상에서의 순위, n은 사용하는 문서구조 정보의 개수이다. 4가지의 구조 정보는 모두 같은 비중을 가진다고 가정하여 전체 구조 정보의 가치는 4가지의 평균값을 이용하여 구한다.

3.3.3. 링크 패턴의 장르 관련값

링크 패턴이 장르에 미치는 영향을 알아보기 위한 전단계로 이미 앞(2장의 장르식별 요소들)에서 장르 문서를 고려한 6가지의 링크 패턴에 대해 살펴보았다. 링크의 수에 따른 링크 패턴들의 영향을 반영하기 위해 수식3을 사용한다.

$$E_k = \frac{\sum_{j=0}^{r} L_j}{\max lf_o} + \frac{\sum_{k=0}^{s} L_k}{\max lf_{in}} \quad (3)$$
$$L = (\text{외부출력, 상위출력, 하위출력, 외부입력, 상위입력, 하위입력})$$

수식에서 L_j, L_k는 링크의 수 L은 가능한 링크 형태, r은 출력 링크 관련 패턴 수, s는 입력 링크 관련 패턴 수를 나타내며 j와 k는 L상에서의 위치를 나타낸다. $maxlf$은 최대링크출현 빈도(max link frequency)를 나타내는데, $maxlf_{out}$, $maxlf_{in}$ 각각은 출력링크와 입력링크에 대한 최대링크출현 빈도를 나타낸다.

1.1.4. 문서 내용의 장르 관련값

내용면에서 장르후보 문서와 사용자가 발견하고자 하는 장르와의 관련성을 알아보기 위한 정보로 활용된다. 초기의 웹 검색 결과에서 확장된 확장 후보문서 집합전체를 하나의 문서 집합으로 가정하여 tf/idf방법을 적용시킨다. 질의 q에 대한 문서 k의 관련도 E_d는 정보 검색에서의 일반적인 유사도 계산 방법[22]을 사용하는데 4와 같이 계산한다.

$$E_{dkq} = \frac{\sum_{i=1}^{n} (t_{ki} \cdot t_{qi})}{\sqrt{\sum_{i=1}^{n} t_{ki}^2 \cdot \sum_{i=1}^{n} t_{qi}^2}} \quad (4)$$

이때 n은 문서 k안의 전체 색인 단어의 개수이며, t는 해당 단어의 가중치를 나타낸다.

1.4. 장르 특성의 융합 및 장르의 판단

각 문서의 장르 관련도를 구하기 위해서 4가지의 장르식별요소들을 사용하였고, 이 증거들을 Dempster-Shaper 방법[6]에 의해 종합적으로 해석하여 문서의 최종 장르 관련도 E_f를 구한다. 방법은 수식 5를 따른다.

$$E_f = \alpha \cdot E_u \oplus \beta \cdot E_t \oplus \gamma \cdot E_s \oplus \delta \cdot E_d \quad (5)$$

수식에서 ⊕는 Dempster-Shaper에서 두 증거의 통합을 위한 연산자에 해당하며, $\alpha,\beta,\gamma,\delta$ 는 각 증거의 반영 비율을 위한 매개 변수로써 실험을 통해 얻어질 수 있는데, 구체적인 값의 결정은 4장에 보인다. E_f는 0에서 1사이의 값을 가지며 1에 가까울수록 높은 장르 관련도를 나타낸다.

4. 실험 및 평가

본 실험의 목표는 각 장르식별 요소들의 역할과 장르 식별 요소들이 장르 발견에 미치는 영향을 알아보는 것이다. 본 실험은 웹 문서를 대상으로 하는데, 아직까지 URL, 링크 정보를 모두 갖춘 웹 문서 실험 집합이 없기 때문에 실험 문서는 일반적인 웹 문서로 정하였다. 발견하고자 하는 장르는 "컴퓨터 분야의 컨퍼런스 홈페이지"로 정하였으며 30개의 컴퓨터 분야를 선정하였다. 장르 발견을 위해 사용할 각 장르에 대한 질의는 표3과 같다.

[표 3] 30개 컴퓨터분야의 장르발견을 위한 질의리스트

Conference Hypertext
Conference Information retrieval
Conference Computer graphics
Conference Artificial intelligence
Conference Algorithms
Conference Databases
Conference Neural networks
Conference Computer networks
Conference Parallel computing
Conference World wide web(www)
Conference Programming languages
Conference Software engineering
Conference Image processing
Conference Data modeling
Conference Java technology
Conference Entity relationship modeling
Conference Fault tolerant computing
Conference High performance computing
Conference Information system design
Conference Conceptual structure
Conference Knowledge and data engineering
Conference Machine learning
Conference Natural language understanding
Conference Network standards
Conference Object oriented computing
Conference Oriental languages
Conference Software reliability
Conference User modeling
Conference VLSI design
Conference Realtime systems

각 질의에 대한 정보발견의 정확도의 평가를 위해서 충남대학교 컴퓨터과학과 대학원생 3명으로 평가 팀을 구성하였는데, 이들은 모두 컴퓨터 분야의 컨퍼런스 홈페이지를 판별할 수 있는 능력을 지니고 있다. 이들 평가자 각각에게는 30개의 장르가 주어지는데, 이들은 본 논문에서 제안하는 방법에 의한 정보발견 결과와 비교 대상 검색 엔진의 결과들에 대해서 장르에 합당한 정보인지를 결정한다. 비교 대상 엔진으로는 웹 검색 엔진인 알타비스타와 메타 검색 엔진인 메타크롤러를 선정하였다. 전자는 제안 알고리즘의 출발점이라는 점에서, 후자는 제안 알고리즘과는 다르지만 여러 웹 검색엔진들의 결과를 이용하여 다른 처리를 한다는 점에서 비교 대상으로서의 의미를 지닌다.

본 실험에서는 최종 판단에 영향을 미치는 장르식별 요소들의 여러 조합과 각각의 반영 비율에 대한 조합을 사용하여 표 4와 같은 실험 결과를 내었다. 총 4개의 장르 식별 요소들(사용 증거)에 대해 한 개에서 네개까지의 조합을 사용하였고 각각의 요소들의 반영 비율을 0.1에서 1.0 까지 0.1 단위로 변화 시켜 나갔다. 최종 판단은 상위 k개에 대한 적합 문서 개수 혹은 정확도를 사용한다. k값은 실험에 의해 최대 30개 단위는 5로 정하였다. 30개를 넘는 지점부터는 제안 알고리즘과 비교 검색 엔진들 모두에서 정확도가 완화되는 경향을 보였으며, 웹 검색 사용자가 평균 상위 30개의 검색결과만을 본다는 점과도 부합된다.

정확도의 계산은 식 6을 따른다.

$$\text{정확도}(\%) = \frac{\text{적합문서의수}}{\text{상위에 대한 발견 문서의 수}} \times 100 \quad (6)$$

표4의 열은 크게 3부분으로 구성 되는데 각 열은 각각 사용 증거들의 조합 개수, 장르 식별 요소의 반영 비율, 그때의 적합문서 개수와 정확도(%)를 나타내고 있다. 예를 들어 7번째 행은 URL이름과 링크 정보의 반영 비율을 각각 0.2와 0.7로 했을 때 상위 5개의 문서에 대해서는 적합 문서 개수가 평균 1.67개이며 정확도는 33.4%임을 나타내다. 마찬가지로 상위 30개에 대해서는 적합문서개수와 정확도가 각각 6.20개와 20.7%임을 보인다.

가장 아래 2행은 비교 대상 엔진으로 선정 된 알타비스타와 메타크롤러에 대한 정확도 실험 결과를 보인다.

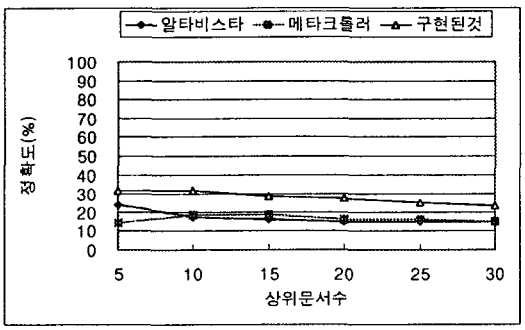
1개의 증거를 사용했을 때, 비교 검색엔진과의 정확도 비교를 통해 4개의 증거는 모두 장르 식별에 도움이 됨을 알 수 있는데, URL 이름, 링크 정보, 구조정보, 문서 정보의 순으로 장르 식별에 높은 영향을 주고 있으며, 각각의 향상 정도는 알타비스타에 대해서는 최대 80%, 메타크롤러에 대해서는 최대 70%의 정확도의 향상을 보인다. 특히, 문서 정보는 확장 후보 집합에 대해 알타비스타와 같은 방식으로 장르 관련도를 계산하는 방법을 취했는데, 알타비스타와 다른 점은 링크 확장을 하였다는 점이다. 이 결과를 통해 링크 확장을 했을 경우 정보 발견의 정확도가 향상됨을 알 수 있었다. 장르 식별 요소의 각 조합을 사용했을 경우 모두 정확도의 향상에 기여를 하고 있으나, 문서 정보는 별 영향을 미치지 못한다는 점도 특이할 사항이다.

[표 4] 장르 발견에 대한 실험 결과

사용증거의 개수	장르식별 요소				적합 문서개수/정확도(%)											
	URL 이름	링크 정보	구조 정보	문서 정보	상위5		상위10		상위15		상위20		상위25		상위30	
1	x				1.13	22.6	2.03	20.3	2.93	19.5	4.17	20.9	4.80	19.2	5.53	18.4
		x			1.23	24.6	2.17	21.7	3.20	21.3	4.00	20.0	4.63	18.5	5.23	17.4
			x		0.93	18.6	1.83	18.3	2.63	17.5	3.73	18.7	4.37	17.5	5.03	16.8
				x	0.83	16.6	1.37	13.7	2.13	14.2	2.80	14.0	3.67	14.7	4.27	14.2
2	0.2	0.7			1.67	33.4	3.13	31.3	4.20	28.0	5.13	25.7	5.80	23.2	6.20	20.7
	0.9		1		1.10	22.0	2.33	23.3	3.07	20.5	4.27	21.4	5.10	20.4	5.80	19.3
	0.2			x	1.13	22.6	2.03	20.3	2.93	19.5	4.17	20.9	4.80	19.2	5.53	18.4
		1	0.1		1.40	28.0	2.57	25.7	3.53	23.5	4.23	21.2	4.87	19.5	5.67	18.9
		0.5		x	1.23	24.6	2.17	21.7	3.20	21.3	4.00	20.0	4.63	18.5	5.23	17.4
			1	x	0.93	18.6	1.83	18.3	2.63	17.5	3.73	18.7	4.37	17.5	5.03	16.8
3	0.3	1	0.1		1.63	32.6	3.10	31.0	4.07	27.1	4.93	24.7	5.90	23.6	6.43	21.4
	0.2	0.7		x	1.67	33.4	3.13	31.3	4.20	28.0	5.13	25.7	5.80	23.2	6.20	20.7
	0.9		1	x	1.10	22.0	2.33	23.3	3.07	20.5	4.27	21.4	5.10	20.4	5.80	19.3
		1	0.1	x	1.40	28.0	2.57	25.7	3.53	23.5	4.23	21.2	4.87	19.5	5.67	18.9
4	0.3	1	0.1	x	1.63	32.6	3.10	31.0	4.07	27.1	4.93	24.7	5.90	23.6	6.43	21.4
알타비스타					0.80	16.0	1.23	12.3	1.63	10.9	2.03	10.2	2.60	10.4	3.07	10.2
메타크롤러					0.63	12.6	1.43	14.3	2.13	14.2	2.50	12.5	3.00	12.0	3.30	11.0

각 조합의 결과를 분석해 볼 때, 장르 식별에 큰 영향을 미치는 순서는 URL이름, 링크 정보, 구조 정보, 문서 정보 순이었으며 앞의 3가지 혹은 4가지 증거를 모두 사용하고 반영 비율을 0.3, 1, 0.1로 했을 때 최대 21.4%의 정확도를 보이고 있다.

결국 장르 식별 요소의 적절한 조합을 사용하여 장르 발견의 정확도를 높일 수 있었다. 그림 5는 제안한 알고리즘과 비교 엔진들의 정확도를 그래프로 보이고 있다.



[그림 5] 각 알고리즘들의 정확도

표5는 제안한 알고리즘의 비교 엔진들에 대한 정확도 향상 정도를 보인다. 결과는 알타비스타와 메타크롤러에 비해 평균적으로 67.34%, 71.78%의 신뢰도 향상을 보임을 입증하였다.

[표 5] 제안 알고리즘의 정확도 향상

상위 문서 수	알고리즘의 정확도(%)			비교대상에 대한 성능향상도(%)	
	제안	알타비스타	메타크롤러	알타비스타	메타크롤러
0	0.00	0.00	0.00	0.00	0.00
5	32.00	24.00	14.60	33.33	119.18
10	32.00	17.30	18.70	84.97	71.12
15	28.87	16.47	18.67	75.30	54.64
20	27.65	15.00	16.35	84.33	69.11
25	25.60	14.92	16.00	71.58	60.00
30	23.33	15.10	14.90	54.53	56.60
평균				67.34	71.78

5. 결론 및 향후 연구 과제

본 논문에서는 정보발견의 일환으로서 웹상에서 텍스트 이외의 URL, 문서 구조, 링크 등의 새로운 식별 요소들을 사용자 중심의 장르발견에 활용함으로써 정보발견의 신뢰도를 향상 시킬 수 있음을 보였다. 또한 각 식별 요소들의 정보발견에 대한 영향력도 알아보았다. 30개의 컴퓨터 분야에 대한 장르발견 결과를 알타비스타, 메타크롤러와 비교해 볼 때, 평균적으로 전자에 대해서는 67.34% 후자에 비해서는 71.78%의 신뢰도 향상을 보였다.

웹의 새로운 식별 요소들을 고려한 장르는 기존의 사용자 프로파일을 이용한 검색을 보완하여 사용자의 정보발견을 도울 수 있다. 여러 분야에 대한 장르 정보를 가진 장르 서버를 구성함으로써 사용자는 자신이 꼭 필요로 하는 장르를 효과적으로 발견할 수 있다. 그러나 장르의 발견을 위해서는 기존 검색에 비해 많은 시간을

필요로 하는데, 주로 링크 확장 및 로봇에 의한 결과 문서 수집에 걸리는 시간으로 이 부분에 대한 향상이 필요하다. 또한 현재는 각 장르식별 요소들이 이진 값을 가지는 경우도 있는데, 장르 식별 요소의 보다 정확한 추출 방법이 필요하며, Dempster-Shaper이외에 이들 요소들을 효과적으로 융합할 수 있는 방법에 대한 연구가 필요하다.

본 연구는 일반적인 장르발견의 시작점으로서의 의미를 가지는데, 현재는 실험모델의 효과를 입증하기 위해서 장르의 범주를 컴퓨터 분야의 컨퍼런스 홈페이지 발견으로 한정하고 있으나 장르 식별 요소의 변경을 통해서 다른 분야로의 확장도 가능함을 명시한다.

참고 문헌

[1] Alvaro E. Monge and Charles P. Elkan, "*The WEBFIND tool for finding scientific papers over the world wide web*", Proceedings of the 3rd International Congress on Computer Science Research, Nov. 27-29, 1996. Tijuana, Baja California, Mexico.

[2] Bowman C Mic, Danzig Peter B, Manber Udi, and Schwartz Michael F. "*Scalable Internet Resource Discovery: Research Problems and Approaches*", University of Colorado Technical Report, 22 Oct 1993.

[3] Brad Perry and Wesley W. Chu, "*Discovering Similar Resources by Content Part-Linking*", Proceedings of CIKM, 1997.

[4] Carriere J. and Kazman R. "*WebQuery: Searching and Visualizing the Web through Connectivity*", Proceedings of the Sixth International World Wide Web Conference, 1997.

[5] Deutsch, Peter. "*Resource Discovery in an Internet Environment*", Master of Science Thesis, McGill University, Montreal, June 1992.

[6] Elaine Rich and Kevin Knight, Artificial Intelligence, 2nd Edition, McGraw-Hill, 1991.

[7] Hermann Kaindl, Stefan Kramer and Luis Miguel Afonso, "*Combining Structure Search and Content Search for the World-Wide-Web*", Proceedings of Hypertext, 1998.

[8] Iannella R, Ward N, Wood A and Leong, D. "*Resource Discovery – A definition*", Distributed System Technology Centre Technical Report.

[9] Jonathan Shakes, Marc Langheinrich and Oren Etzioni, "*Dynamic Reference Sifting: A Case Study in the Homepage Domain*", Proceedings of 6th World Wide Web, 1997.

[10] Kleinberg, J. "*Authoritative sources in a hyperlinked environment*", Proc. of 9th ACM SIAM symposium on Discrete Algorithms. Also appeared as IBM Research Report RJ 10076, May 1997.

[11] Krishna Bharat and Monika R. Henzinger, "*Improved Algorithms for Topic Distillation in a Hyperlinked Environment*", Proceedings of SIGIR, 1998.

[12] Oren Etzioni and Daniel Weld, "*A Softbot-Based Interface to the Internet*", Comminication of ACM, July. 1994.

[13] Robin Burke, Kristian Hammond and Julia Kozlovsky, "*Knowledge-based Information Retrieval from Semi-Structured Text*", Proceedings of AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, pp 19-24, 1995.

[14] Search Engine Watch, "Search Engines Sizes", http://www.searchenginewatch.com/reports/sizes.html.

[15] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceedings of the 7th International World Wide Web Conference, 1998.

[16] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson and Jon Klenberg, "*Automatic resource compilation by analyzing hyperlink structure and associated text*", Proceedings of the seventh International World Wide Web Conference, 1998.

[17] William W. Cohen. "*Knowledge Integration for Structured Information Sources Containing Text(Extended Abstract)*", Proceedings of SIGIR Workshop on Networked Information Retrieval. 1997.

[18] WonKyun Joo and SungHyon Myaeng, "*Improving Retrieval Effectiveness with Link Information*", Proceeding of the 3rd International Workshop on Information Retrieval with Asian Languages(IRAL'98), October. 15-16, 1998. Singapore.

[19] Yeong, Wengyik. "*Towards Networked Information Retrieval*", Technical Report 91-06-25-01, Perfomance Systems International, Virginia, USA. 1991.

[20] Marchioni M. "*The Quest for correct Information on the Web: Hyper search Engines*", Proceedings of the Sixth International World-Wide Web Conference, 1997.

[21] 주원균, 맹성현, "*디지털 도서관을 위한 로봇 에이전트의 설계 및 구현*". 한국정보과학회 가을 학술발표 논문집, Vol.25, No.2, pp.433~435, 1998.

[22] Gerard Salton, Automatic Text Processing, Addison-Wesley Publishing Company, 1989.

[23] 임정묵, 오효정, 맹성현, 이만호 "*카테고리 정보 활용을 통한 링크 기반 검색의 속도 향상*". 한국정보과학회 봄 학술대회 논문집, 4월, 1999.