제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

웹 상의 정보검색을 위한 지능형 검색시스템의 연구

박병율, 임종태
공주대학교 컴퓨터공학과
e-mail:brpark98@kongju.ac.kr

A Study of Practical Search System for Information Retrieval on the Web

Beung-Raul Park, Jong-Tae Lim
Dept. of Computer Engineering, Kong-Ju University

요 약

검색시스템은 분류시스템과 지식탐사 시스템을 결합하여 구성한 복합적인 시스템으로 일반 사용자들에게 자신이 원하는 정보의 데이터를 우선적으로 제공한다. 시스템의 특징으로 겉으로 보기에는 일반 검색엔진과 유사하나, 시스템적으로는 요구하는 각종 기능과 검색 기법, 지식탐사기법이 들어있다. 시스템에서는 문서 분류기법과 문서와 검색어 사이의 연관성을 찾기 위한 방법, 문서간의 연속적인 사건을 통한 검색 패턴 탐사기법을 사용하였다. 이들은 시스템의 검색과 분류 결과를 지금까지보다 더욱 인공지능에 가깝도록 하여 준다.

1. 서론

현재 인터넷 환경에서 정확한 정보의 전달은 매우 중요한 과제 중의 하나이다. 인류는 사용자가 원하는 정보와 관련 정보의 링크 제공을 위해 많은 연구와 방법을 연구하였으며, 데이터마이닝 및 정보검색 분야에 많은 발전을 하였다.

본 연구는 정보 분류의 정확성과 제공되는 관련 정보의 정확성 향상에 목적을 둔다. 우선순위와 예약어 DB를 이용한 메타데이터의 디렉토리별 분류를 사용자들의 관심도와 시간정보를 기초로 재분류함으로 기존 분류들의 정확도를 높이며, 사용자들의 반응을 이용하여 사용자의 검색패턴을 발견하여 사용자가 원하는 정보를 보다 찾기 쉽도록 도와준다. 정보검색 기술로는 2장에서 소개되는 분류기법들 중 벡터 스페이스모델과 확률모델이 이용하는 검색어의 존재에 대한 가중치를 기반으로 분류하며, 지식 탐사기법을 이용 이를 재분류하는 방법을 택한다. 이런 지식탐사 기법 중 연관규칙과 연속규칙을 적용한 학습원리를 검색엔진에 적용하고 결과를 살펴본다. 사용자 정보를 이용하는 학습 알고리즘의 효과는 문서에 포함된 정확한 정보를 사용자에게 제공하며, 관련 문서에 대한 정보도 함께 제공하여 준다.

2. 분류기법과 지식탐사기법

2.1 분류기법

현재 문서 분류기법을 크게 분류하면, 부울대수를 기반으로 문서와 검색어 간의 관계를 정의하는 Boolean 모델과 n-차원 공간상의 한 점으로 표현하는 벡터 스페이스 모델, 추출된 확률의 적합성을 기반으로 표현하는 확률모델 등 두 가지로 분류할 수 있다. 벡터스페이스 모델은 문서를 n-차원 공간상의 한 점으로 표현하는 방식으로 n은 전체 데이터 셋에서 사용하는 단어의 개수를 의미한다. 각 차원은 단어의 가중치를 나타내며, 가중치의 측정을 위해 문서의 유사성을 측정하고, 유사성이 높은 순서대로 문서를 제공하는 방식이다. 확률모델기법은 추출된 문서들을 적합할 확률에 따라 정렬한다. 확률모델에서 추출된 문서들의 실제 확률은 중요치 않으며, 확률에 따른 문서의 순위가 중요하게 작용한다.

이 두 가지 방식은 문서 내에 단어의 존재 여부를 통해 문서를 분류한다는 공통점을 지니고 있다. 이는 문서와 검색어는 상당한 상관 관계가 있음을 의미한다. 그러나 이 두 가지 방식 모두 단어의 문서 내의 의미는 해석하지 못한 것을 알 수 있다.

2.2. 지식탐사기법

데이터마이닝의 기법 중 가장 널리 사용되고 있는 기법중의 하나가 연관규칙 탐사기법이다. 연관규칙이란 동시 혹은 연속으로 발생하는 두 개 혹은 그 이상의 사건들 사이의 상관관계를 의미한다. A-Priori 알고리즘은 가장 유력한 알고리즘 중의 하나로 필터링 조건 보다 적은 값들이 bound 될 때, 매개변수의 값이나 그 밖의 값들을 고려하여 제거한다. 이러한 연관규칙 탐사를 통해 항목별 연관성을 찾게 된다.

연속규칙 탐사기법이란 특정 항목들이 일련의 순서에 의해 나타나는 경향을 찾고 분석하는 기술을 말한다. 즉 연관성이 있는 항목집합에 연관성과 시간과의 관계를 추가한 것으로, 고장진단 및 고객의 구매 성향분석 등에 이용된다. 또한 이 연관성은 항목집합과 문서와의 상관관계를 이용한 디렉터리의 재분류와 사용자의 문서이용정보를 통해 문서의 상관관계를 분석하는 자료로도 활용된다.

3. 정보분류와 지식탐사

3.1. 정보분류

분류시스템 중 벡터스페이스 모델은 문서 내에 예약어가 존재하는가의 여부와 존재하는 횟수를 기준으로 문서와 예약어 사이의 유사성을 측정한다. 확률 모델의 경우는 문서의 접합성은 예약어의 가중치를 기준으로 한다. 이 둘의 동일한 점은 문서 내에 예약어가 존재하는가의 여부가 관심의 대상이라는 점이다. 예약어의 존재의 유무는 문서의 예약어와의 연관성을 찾는데 이는 매우 중요한 작업이다.

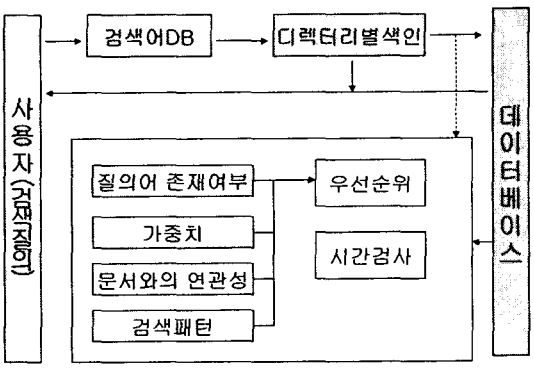

그림1) 지능형 검색시스템의 구성도

그러나 동일한 단어가 두 개 이상의 뜻을 내포하는 경우, 예약어의 존재의 유무만으로는 문서와의 연관성을 구별하기 어렵다. 또한 예약어가 존재하지 않는 경우일지라도 문서와의 연관성을 가지고 있을 경우가 있다. 우리가 단지 "색감"이라는 단어만을 입력하였다고 하자. "색감"이라는 단어에는 '색의 느낌', '색을 내기 위한 재료'라는 뜻이 내포되어 있다. 이런 언어의 다양성을 한두 가지의 방법으로 컴퓨터에 모두 입력할 수는 없다. 문서 분류에 있어 최상의 분류는 인간의 힘으로 분류하는 방법이다. 차선의 방법으로 벡터스페이스 모델과 확률 모델에 의해 문서를 분류한다. 이를 기준으로 문서를 디렉토리별로 분류한다. 이런 분류를 위해 예약어 사전이 필요하며, 정확한 예약어는 유사성과 적합성에 의한 문서 분류의 정확성을 위한 기준이 된다

3.2 지식탐사

앞 절의 방법에 의해 디렉토리별로 분류된 메타데이터 만으로도 일반적인 검색 지원은 가능하다. 그러나 좀 더 나은 검색결과를 제공하기 위해서는 인간의 언어적인 면을 문서분류 안에 포함시키는 작업이 필요하다. 이러한 작업은 연관 규칙을 위한 A-Priori 알고리즘으로 해결할 수 있다. 다만 정확한 문서와 예약어와의 연관성을 모색하기 위한 시간이 필요할 뿐이다.

연관 규칙을 발견하기 위해 사용자들의 특정 문서에 대한 접근 정보를 카운트하여야 한다. 카운트 정보는 사용자들이 기대하는 문서에 대한 정보이다. 사용자의 예약어 그룹에 대한 사용자의 접근 정보(U_i)와 특정 문서의 클릭 횟수(CL_i)와의 관계에서 상관계수 는 $c = CL_i / U_i$ (단 $c \le 1$)로 나타난다. 메타데이터의 생성 일자를 기초로 한 문서의 생명지수 t는

$$t = \frac{Cr_{date} - Ur_{date}}{30day} \quad (\text{단 } t \ge 1)$$

이며, 생명지수가 클수록 오래된 문서가 된다. Cr_{date}는 현재의 년월일을 일로 환산한 값이고, Ur_{date}는 메타데이터가 기록된 년월일을 일로 환산한 값을 의미한다. 생명지수는 우선 순위와 반비례하며, 생명지수의 값이 클수록 문서의 우선 순위는 낮아진다. 상관계수와 생명지수의 값을 근거로 우선순위(k)를 구하면,

$$k = \frac{c}{t} K \ (K \text{ 는 SimF와 W의 산술평균})$$

이 된다. 우선순위의 값은 상관계수가 높은 문서일지라도 오래된 문서의 경우는 매우 작은 값으로 나타나게 된다. 이는 상관계수와 함께 생명지수를 중요시하였기 때문이다. 이는 매 시각별로 급변하는 웹 문서에서 상관계수가 높은 문서이지만 시간에 따라 문서의 중요도가 낮아지게 되기 때문이다. 검색

된 웹 문서는 정보이므로 문서에서 시간에 따른 문서의 재배치는 반드시 필요하다. 우선순위에 의해 분류된 메타데이터는 각 문서에 대해 생명지수와 상관계수 값에 의해 분류됨으로 스스로 학습하는 효과를 갖게 된다. 운용하는 방법에 따라 차이가 있을 수 있으나, 우선 순위의 값이 불능이 되는 문서, 즉 상관계수의 값이 0인 상태로 생명지수의 값이 계속해서 증가되는 문서의 경우는 다른 디렉토리로 이동하여 서비스하는 방법으로 검색된 웹 문서의 사장을 막는 효과를 볼 수 있다.

4. 지능형 정보검색의 구현

4.1. 정보분류시스템

웹 상의 정보검색을 위한 정보분류시스템은 다음과 같은 원칙을 기준으로 한다.

가. 사용자에 의한 우선순위 결정.
나. 시간 정보에 의한 재분류.
다. 문서와 검색어간의 연관성.

먼저 '사용자에 의한 우선순위 결정'은 정보의 우선 순위가 관리자뿐 아니라 사용자의 입력의 유무에 따라 변할 수 있어야 함을 의미한다. 사용자 클릭은 문서내의 예약어의 의미를 파악하는 중요한 정보로서 클릭의 횟수에 따라 우선순위는 변하게 된다. 둘째 '시간 정보에 의한 재분류'는 문서의 시간에 따른 정보로서의 가치를 평가하는 항목이다. 정보란 제시되는 시간이 중요하다. 오늘의 정보가 내일은 상식으로 그치는 경우도 있다. 따라서 정보가 지니는 가치를 시간에 맞게 조절하게 함으로 정보로서의 가치를 높게 한다. 셋째 문서의 우선순위를 유동성이 있는 값들을 기준으로 한다면 우선 순위도 자연 유동성을 갖게 될 것이나 문서 고유의 우선순위는 반영되지 않을 수도 있다. 따라서 우선순위에는 변하지 않는 값들인 예약어의 유사도, 적합도 및 사용자 입력 정보인 상관계수와 생명지수를 포함한 값이 되도

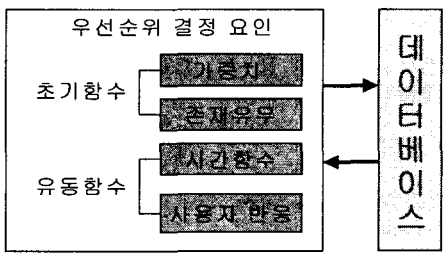

그림2) 우선순위를 결정할 수 있는 요인들

록 한다.

지금까지의 결과를 기준으로 문서의 우선순위를 정하고 이후는 사용자의 선택에 따른 정보를 통해 우선순위를 재 선정하도록 한다. 사용자의 정보선택은 다른 우선순위의 기준과 비교하여 조금도 손색없는 기준이 된다. 사용자의 선택이 많은 정보는 그만큼 효용가치가 있다고 평가할 수 있다. 따라서 다음 절에서 사용자의 선택을 우선순위에 포함시키는 방안을 살펴본다.

4.2 지식탐사 시스템

일반적으로 사용자가 질의어를 입력하면 시스템은 질의어가 원하는 문서를 찾기 위하여 메타데이터를 검색하여 질의어와 일치하거나 관련이 있는 문서를 검출함으로 사용자의 질의에 응답하게 된다. 이 경우 사용자의 반응은 철저히 배제되며, 사용자의 응답은 시스템과 무관하다.

본 연구의 시스템은 이러한 정보검색 기법에 약간의 수정을 가한다. 분류기법을 통해 디렉토리 별로 분류된 정보에 사용자의 선택을 추가하여 좀 더 발전된 정보를 제공하는 방법을 선택한다. 예를 들어 사용자가 '포트리스'라는 단어를 질의어로 입력하였을 경우, 시스템은 '컴퓨터'→'게임', '오락/게임'→'컴퓨터게임', '엔터테인먼트'→'영화' 디렉토리로 이동하여 관련 항목집합을 찾는다. 검색된 결과는 화면을 통해 출력되며, 사용자는 자신이 원하는 문서를 클릭함으로 내용을 검색한다. 이때 사용자 선택을 화면을 통해 문서의 우선순위 변경을 할 수 있도록 데이터베이스에 저장시킨다. 축적된 사용자 선택은 시간에 따른 사용자의 문서에 대한 선호도를 나타내는 지수의 역할을 하게되며 시스템은 이것을 평가하며 문서와의 연관성을 재평가하게 된다.

ID	Time	Items
1	02/08/02	2, 5, 8
1	02/08/03	8, 2, 5
1	02/08/10	8, 3, 5, 2
1	02/08/11	3, 4, 8
2	02/08/05	7, 5, 9
2	02/08/07	3, 5, 7
2	02/08/12	4, 5, 9
3	02/08/03	3, 7, 9
3	02/08/08	7, 8, 9
3	02/08/11	5, 8, 9
4	02/08/10	3, 9
4	02/08/13	5, 7, 9

ID	Sequence
1	<2,5,8> <3, 8>
2	<5,7> <5, 9>
3	<7,9> <8, 9>
4	--

검출된 연속패턴 중 지지율 70%이상인 패턴 :

1: < 2, 5, 8>

그림3) 연속규칙을 통한 검색패턴의 발견

공인된 사용자들의 경우, 시스템으로 사용자의 검색패턴을 과거의 정보 선택을 기준으로 판단하여 정보서비스를 함으로 맞춤 정보를 제공받는다. 이 경우 질의어의 입력 정보는 자신만을 위한 공간에 저장되며, 자신의 정보활용의 패턴을 검출하는 기준으

로 활용되게 한다.

5. 결론

문서를 분류하는 방법으로는 많은 분류 방법이 있다. 본 연구는 벡터 스페이스 모델과 확률 모델을 기본적인 근간으로 한다. 시스템은 이들 분류 모델이 가지지 못한 질의어의 의미를 분석하는 기준을 찾는데 목적이 있었다. 본 논문은 질의어의 의미를 판단하는 방안으로 데이터 마이닝 기법 중 연관규칙과 연속규칙 기법을 사용하였다. A-Priori 알고리즘에 기반한 연관 항목의 발견은 사용자들의 선택을 근거로 문서의 우선 순위를 결정하는 역할을 한다.

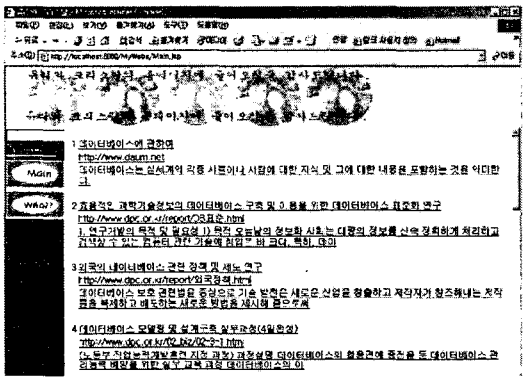

그림4) 지능형 검색시스템의 예

물론 사용자의 선택에 대한 정확성의 검증 문제도 있으나, 다운로드나 문서를 직접 여는 경우만을 카운트함으로 오차의 폭을 최대한 줄이고자 노력하였다. 또한 공인된 사용자들을 위한 검색패턴을 찾는 방안은 앞으로 맞춤 검색의 시대를 열 수 있을 것으로 기대된다.

본 논문은 전문가의 힘을 빌리지 않고 시간의 흐름에 따라 문서를 정확히 분류하는 방안을 제시하였다. 앞으로 본 연구는 컴퓨터가 인간의 힘에 의지하지 않고 자신의 힘으로 정보를 정확히 분류하는 방안을 찾는데 더 많은 중점을 둘 것이다. 다만 기반이 되는 지식의 습득 내지 학습을 위해 필요한 시간은 입력 및 기존의 정보를 활용함으로 줄이게 된다. 끝으로 정보검색과 데이터마이닝을 통한 지식 발견에 괄목할 만한 성과들을 기대해 본다.

참고문헌

[1] 장병탁 외4, "지능형 인터넷정보 서비스를 위한 대규모 텍스트 분류 및 검색기술에 관한 연구", 2001년도 대학기초사업 최종 연구개발결과보고서 서울대학교 2001. 7. 31.

[2] 이상조 외6, "Web상에서 정확한 검색을 위한 문서의 대표개념어 생성 및 요약 시스템에 관한 연구", 2001년도 대학기초사업 최종 연구개발결과보고서 경북대학교 2001. 7. 14.

[3] 윤종필외 2 "데이타 마이닝의 유용성", 정보과학회지 제16권 제8호, pp. 16~20, 1998. 8.

[4] 이도헌외1 "데이타 마이닝 기술 및 연구동향", 정보과학회지 제16권 제9호, pp. 6~14, 1998.9.

[5] 지원철외1 "데이타 마이닝과 의사결정 지원 시스템", 정보과학회지 제16권 제9호, pp.24~36, 1998.9.

[6] 최종후외3 "Answer Tree를 이용한 데이터마이닝 의사결정나무분석", SPSS아카데미. pp. 19~72

[7] B.Y Ricardo and R. N Berther "Modern Information Retireval" ACM Press

[8] M. S. Chen, J. Han, and Philip S. Yu, "Data Mining: An Overview from Database Perspective", IEEE Trans. on Knowledge and Data Engineering, 1997.

[9] T. Dick, Jeffrey D. Ullman, and et al., "Query Flocks: A Generalization of Association- Rule Mining", ACM SIGMOD, pp. 1~12, 1998,

[10] T. Fukuda, Y. Morimoto, S. Morishta, and T. Tokuyama, "Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization", ACM SIGMOD, pp.13~23, 1996.

[11] M. Gilman, "Nuggests and Data Mining", White Paper http://www.data-mine.com/

[12] V. Gudivada, V. Raghavan, W. Grosky, and R. Kasanagottu, "Information retrival on World Wide Web.", IEEE Internet Computing, pp.58~68 Sept~Oct. 1997.

[13] J. Han, Y. Fu, and W. Wang, "DBMiner: A System for Mining Knowledge in Large Relational Databases", Proc. Int'l conf. on Data Mining and Knowledge Discovery(KDD 96), pp.250~255, 1996.

[14] E. H. Han, G. Karypis, and V. Kumar, "Scalable Parallel Data Mining for Association Rules.", ACM SIGMOD, pp.277~288, 1997.

[15] J. Han, "Towards On-Line Analytical Mining in Large Databases", ACM SIGMOD Record, pp. 97~107, March 1998.

[16] Y. Kotidis and N. Roussopoulos, " An Alternative Storage Organization for ROLAP Aggregate Views Based on Cubetrees", ACM SIGMOD, pp.249~258, 1998.

[17] J. S. Park, M.S. Chen, and P. S. Yu, "An Efficient Hash-Based Algorithm for Mining Association Rules", ACM SIGMOD, pp.175~186, 1995.

[18] J. S. Park, M. S. Chen, and P. S. Yu, "Data Mining for path Traversal patterns in a Web Environment", Proc. the 16th ICDCS, pp 385~392, 1996.