웹과 GIS를 통합한 "Kyonggi21Search" 구현 : 색인어간 연관도 생성 및 최적화

장정훈○*, 이룡**, 上林彌彦**, 권용진*
*한국항공대학교 정보통신공학과, **일본교토대학 대학원 사회정보학전공
jhjang@tikwon.hangkong.ac.kr

Implementation of "Kyonggi21Search" combining GIS with The Web : Optimization of Index Association

Jung-Hoon Jang*, Ryong Lee**, Yahiko Kambayashi**, Yong-Jin Kwon*
*Dept. of Telecommunication and Information Eng. of Hankuk Aviation University
**Dept. of Social Informatics, Graduate School of Informatics, Kyoto University

요 약

"Kyonggi21Search"시스템은 GIS와 웹을 통합한 지역정보 검색 시스템이다. 웹과 GIS를 연동하여 지리정보를 검색하기 위해 웹 문서에서 지역관련 색인어를 추출하고, 색인어간의 관련성을 계산한다. "Kyonggi21Search"시스템에서는 웹 문서에 많이 나타나는 일반적인 단어보다는, 많은 문서에 나타나지 않는 지리적, 문화적인 단어들 간의 관련성을 찾는 것이 더 중요한데, 본 연구에서는 단어들 간의 관련성을 찾는데 연관규칙과 연관클러스터를 이용하여 연관도를 계산한다. 그리고 이런 단어들의 관련성을 찾는데는 연관 클러스터를 이용하는 것이 더 적합하다는 것을 보여준다. 한편 웹 문서와 색인어를 이용하여 만든 행렬은 희소행렬이라는 점을 이용하여 연관 클러스터 방법의 단점인 높은 계산량을 줄이는 최적화 방법을 제안한다.

1. 서 론

웹은 전례 없는 규모로 생각과 정보의 공유를 가능하게 하였으며, 인류 지식과 문화의 보편적인 저장소가 되었다. 또한 각종 정보는 그 양적인 측면에서 폭발적으로 증가하고 있는데, 이러한 웹의 방대함은 어떤 사용자라도 자기 자신의 웹 문서를 만들 수 있으며, 제한 없이 다른 웹 문서를 연결 할 수 있게 되었기 때문이다.[1]

이러한 웹의 특성으로 정보의 양은 많아졌지만, 웹 상에서 유용한 정보를 찾는다는 것은 어려운 일이 되어 버렸다. 이런 어려움에 대한 해결책으로 정보 검색과 그 기술에 대한 관심이 증대되고 있다. 누구나 웹의 문서를 만들 수 있다는 웹의 특성상 정보의 양은 많아지지만 정보의 정의와 구조가 저 수준이 되므로, 항상 좀 더 효율적인 정보 검색 시스템에 대한 요구가 증대되고 있다.

이런 요구에 대한 검색 시스템의 하나로써 웹에서 이용 가능한 지리정보 시스템이 있다. 지리정보 시스템이란 스캐닝 되었거나 디지털 형태로 변환된 지형 데이터를 이용하여 지도를 보여(display)주는 시스템을 말하는데, 많은 종류의 데이터들이 중요한 지리적 형상을 가지고 있기 때문에 그 목적에 맞게 여러 가지 형태로 데이터를 가공하여 필요한 부분만 중점적으로 보여질 수 있다. 이러한 특성으로 교통, 일기예보, 인구예측, 문화재 관리, 토지계획 및 여행지의 위치정보 등 다양한 분야에 응용되고 있다[2]. 웹 정보공간과 현실공간과의 관련짓기(mapping)를 목표로 하는 연구는 많이 있지만, 대부분은 웹 정보를 얼마나 효율적으로 현실 공간의 장소에 매핑 할지, 또는 지리정보 시스템 적인 발상에서 현실공간거리를 고려한 웹 정보 검색에만 집중되고 있는 현실이다. 지리정보 시스템이 웹에서 서비스를 하고 있음에도 불구하고 웹 정보의 이용은 아직 초보적인 단계이다.

"Kyonggi21Search"시스템 구현을 위한 연구에는 웹 문서 수집, 키워드 집합 구성을 위한 색인어 추출, 키워드간의 연관도 분석 등이 있다.

본 논문에서는 단어간 연관성을 분석하여, 단순한 지리 정보뿐만 아니라 그와 관련된 지역 지식을 함께 보여주는 "Kyonggi21Search"시스템[3]의 키워드 생성에 적합한 연관도를 생성하고, 최적화시키는 방법을 제시한다.

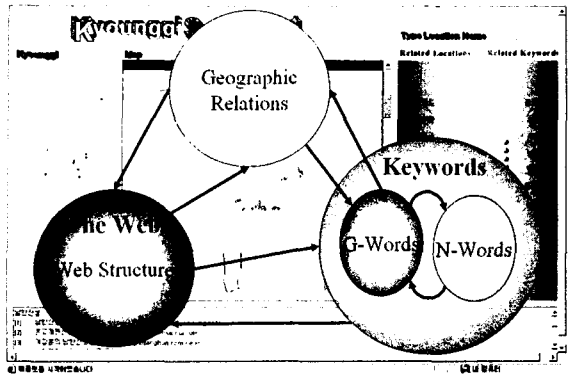
본 논문의 구성은 다음과 같다. 제2장에서는 단어간 연관성을 분석하는 기존의 관련연구에 대해 기술하며, 제3장에서는 기존의 방법들을 "Kyonggi21Search"시스템에 적용하여 적합한 연관도 생성방법을 제시하고, 제4장에서는 제시된 방법의 최적화에 대해 제안한다. 마지막으로 제5장에서는 결론과 향후 연구과제를 기술한다.

2. 관련연구

"Kyonggi21Search"시스템은 지도와 웹을 동시에 검색하여 지리정보를 효율적으로 찾아 볼 수 있는 시스템이다. 인터페이스는 <그림1>과 같이 지도인터페이스, 키워드인터페이스, 웹페이지인터페이스로 구성되어 있다. 지도와 웹을 연동하기 위해 웹 문서에서 지역관련 색인어를 추출하여 키워드 검색이 가능하도록 하였는데, 이때 키워드를 가장 관련이 높은 키워드와 연결되도록 키워드간의 연관도를 생성한다. 키워드간의 연관도

† 본 논문은 과학기술부·한국과학재단지정 「한국항공대학교 인터넷정보검색연구센터」의 연구비 지원으로 수행되었음.

를 생성하는 방법에는 여러 가지가 있는데, 연관규칙(Association Rule)을 찾는 apriori 알고리즘과 행렬의 곱을 이용하는 연관 클러스터의 두 가지 방법을 살펴본다.



<그림 1 Kyonggi21Search의 인터페이스>

2.1 연관규칙(Association Rule) 생성[6]

연관규칙은 두 색인어 집합을 동시에 포함하는 문서의 수를 전체 문서의 수로 나누어준 지지도(Support)와 두 색인어 집합을 동시에 포함하는 문서의 수를 한 색인어 집합을 포함하는 문서의 수로 나누어준 신뢰도(Confidence)라는 두 척도를 이용하여 색인어간 연관성을 찾아내는 것이다[4][5]. 동일한 웹 문서에 나타나는 색인어들은 "한번에 함께 산 물건들"로 볼 수 있으며, 동일한 문서에 나타난 색인어들은 서로 관련이 있다고 판단한다. 만약 어떤 색인어들이 동시에 많은 문서에 출현한다면 그 색인어들은 연관성이 높다고 할 수 있다.

연관규칙을 찾는 대표적인 알고리즘에는 apriori 알고리즘이 있다. apriori 알고리즘은 다음과 같이 두가지 부문제로 규칙을 찾아낸다.

① 최소지지도(minsupport) 이상이 되는 항목들의 모든 조합을 찾는다. 이것들을 빈발항목집합(large itemsets)라 부른다.

② 앞에서 얻어진 빈발항목집합을 $Y = I_1, I_2, \cdots, I_k$ 하고 규칙들의 선항을 Y의 부분집합 X라 하면, 규칙 $X \Rightarrow I_j | c$를 생성하기 위해 X의 지지도로 Y의 지지도를 나눠준다. 이 값이 신뢰도 c보다 크면 $X \Rightarrow I_j | c$이 규칙이 된다.

위의 빈발항목집합을 구하는 문제는 "어떤 집합이 주어졌을 때, 새로운 항목을 더해주면 지지도는 절대로 전보다 증가 할 수 없다"는 것에 착안한다. AB가 빈발항목집합이 아니라면, ABC, ABD와 같은 항목집합들은 절대로 빈발항목집합이 될 수 없으므로, AB가 들어가는 조합은 제외시키고 빈발항목집합을 구한다.

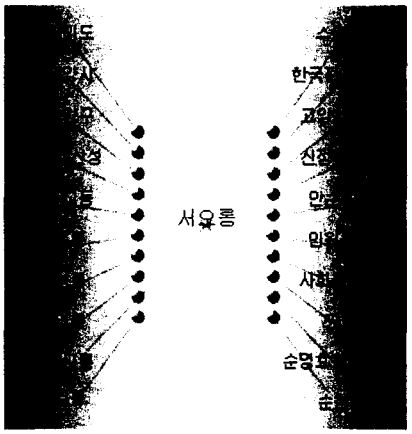
2.2 연관 클러스터를 이용한 색인어간 연관도 생성

연관 클러스터는 단어들의 문서 내 공기 관계를 이용한다. 우선 문서 d_j에 색인어 w_i가 몇 번 출현하는지 빈도를 측정하여 행렬 m_{ij}를 만들고, m_{ij}의 전치행렬 m'_{ij}를 만들어 두 행렬의 곱 $s = m_{ij} \times m'_{ij}$를 만든다. 행렬 s의 요소는 두 색인어간의 연관도가 된다[1].

두 색인어가 동시에 많은 문서에 나타날수록 두 색인어간 연관도가 높아지게 되는 것은 연관규칙과 같다. 다른 점은 각 문서에 나타나는 횟수도 연관성에 기여하므로 여러 문서에 나타나지 않더라도 한 문서에 많이 나타나면 강한 연관성을 갖게 된다는 점이다.

3. 연관규칙과 연관클러스터의 비교

"Kyonggi21Search"시스템에서 사용자가 색인어를 입력하면, 사전에 분류된 "경기도" 관련 색인어 중에서 연관도가 높은 20개의 단어를 뽑아서 보여준다. 예를 들어 사용자가 "서오릉"이라는 단어를 키워드로 선택하여 텍스트 검색 창에 입력하면 "경기도", "고양시", "행주산성", "홍릉"과 같은 단어들이 새로운 키워드로써 20개가 사용자에게 제시된다. 이렇게 제시된 키워드 중 "행주산성"에 흥미가 있는 사용자는 "행주산성"을 새로운 키워드로 사용할 수 있고, 역시 "행주산성"에 대해 20개의 새로운 단어들이 생성된다. 이와 같은 과정들은 "Kyonggi21Search"시스템에서 <그림 2>와 같은 인터페이스로 나타난다.



<그림 2 Kyonggi21 Search의 키워드인터페이스>

본 연구에서는 색인어간 연관성을 찾아내는 기술로, 데이터 마이닝의 연관규칙(Association Rule)과 질의 확장에서 쓰이는 연관 클러스터(Association Cluster) 방법을 사용하여 두 결과를 비교한다.

3.1 연관규칙을 이용한 색인어간 연관도 분석

일반적인 연관규칙의 알고리즘으로 찾아내는 연관성은 $\{w_1, w_2, \cdots, w_{k-1}\} \rightarrow \{w_k\}$ 같이 원소의 개수가 k인 집합의 모든 부분집합에 대해 연관성이 강한 w_k를 찾는 것인데[8], 본 연구에서 필요한 두 단어간의 연관성은 k가 2인 경우에만 생각하면 된다.

먼저 최소지지도와 최소신뢰도가 각각 0.04%, 7% 수준에서 얻어진 연관규칙에서 "행주산성"과 연관도가 높다고 나온 단어들은 "고양시", "일산구", "덕양구" ... "행주대첩비", "행주대첩제", "행주문화제"등 73개가 선택되었다. 본 연구에서 필요한 키워드의 개수는 20개로 한정하였으므로, 이 단어들 중 연관도가 높은 단어 20개를 선택해야 한다. Apriori에서 더욱 연관도가 강한 단어를 추출하기 위해서는 최소지지도와 최소신뢰도를 높여주면 된다. 각각 0.05%, 8% 수준에서 얻어진 연관규칙에서 "행주산성"과 연관도가 높다고 나온 단어들은 총 28개가 선택되었다. 이와 같은 방법으로 "Kyonggi21Search"시스템에서 필요로 하는 키워드를 추출할 수 있다.

하지만, 두 번째 실험에서 "행주산성"과 연관성이 있다고 선택된 단어 28개 중 경험적 판단에 의해 연관성이 높다고 할 수

있는 "행주대첩비", "행주대첩제", "행주문화제"와 같은 단어들이 탈락되었다. 이것은 "행주산성"과 이 단어들을 동시에 포함하는 문서가 전체 6825개 문서 중 3개 밖에 없기 때문에 최소 지지도 0.05%에서는 탈락 될 수밖에 없다. 이와 같이 지지도와 신뢰도를 이용하여 연관단어를 얻어내는 연관규칙은 한 문서에 색인어가 몇 회 출현하는지 상관없이 1회로 결정하여 연관도를 계산하기 때문에 "행주산성"과 경험적으로는 연관도가 높은 단어가 극소의 문서에만 나타나서 탈락된다.

3.2 연관 클러스터를 이용한 색인어간 연관도 분석

이 방법을 이용하여 "행주산성"과 연관도가 높다고 얻어진 단어들은 "고양시", "북한산성" ... "행주대첩비"등인데, 이 결과집합이 Apriori를 이용한 방법과 대체로 비슷하지만, "행주대첩비"와 같은 단어가 연관도 순으로 15위에 랭크되어 있다. 이것은 대체로 여러 문서에 같이 존재하는 단어일수록, 특정문서에 많은 빈도로 같이 나타난 단어일수록, 연관도가 높게 평가되기 때문이다. 실제로 "행주산성"을 포함하는 문서의 수는 42개이고, 이 문서 중 "행주산성"이라는 단어가 15회, 11회의 고빈도로 출현하는 문서가 두 개 존재하는데, 이 문서들에는 "행주대첩비"라는 단어가 각각 1회, 4회씩 출현하고 있다. "행주산성"과 "행주대첩비"가 같은 문서에서 각각 11회, 4회의 고빈도로 출현하기 때문에 이 때의 특성이 반영이 되어서 연관 클러스터 방법에서는 "행주대첩비"가 "행주산성"과 연관 단어로써 선택된 것이다.

위와 같은 특징 때문에 "경기도"와 같이 전체적으로 많은 문서에 출현하는 단어의 경우는 특정 문서에서 어떤 단어가 어느 정도의 빈도로 나타났느냐 하는 특성보다는 여러 문서에 같이 출현하는 단어가 연관성이 높게 나오므로 Apriori를 이용한 방법과 큰 차이를 보이지 않는다. 하지만, "행주산성"과 같이 지리적, 문화적인 특성을 갖는 단어인 경우는 "Kyonggi21 Search"시스템에서는 꽤 중요한 단어인데, 이런 단어들은 전체적으로 많은 문서에 출현하지 않기 때문에 이 단어들과 연관성이 높은 단어를 추출하는 데는 연관 클러스터 방법이 좀 더 효과적이라고 할 수 있다.

4. 희소행렬을 이용한 연관 클러스터의 최적화

색인어간 연관도 분석에서 두 행렬의 곱을 이용하는 방법은 막대한 메모리와 계산량을 필요로 한다. 고양시 관련 단어 1945개와 고양시 관련 웹 문서 6825개로 연관 클러스터 방법을 이용 할 때는 많은 메모리를 필요로 하지 않는다. 하지만, 새로 추출한 경기도 관련 단어 23,238개와 웹 문서 119,459개로 연관 클러스터 방법을 이용하려면 행렬 m_{ij}을 만드는 것만으로도 약 2.7기가바이트의 메모리가 필요하다. 여기에 웹 문서를 더 모으고 관련 단어를 좀 더 추출한다면 행렬 m_{ij}를 표현하기는 더욱 어려워진다. 한편, 행렬 s의 각 요소 하나를 만드는데 곱셈과 덧셈 연산의 총 횟수는 230,000회이며, 행렬 s의 성분은 약 2,700,000,000개만큼 있다. 이 행렬을 올려놓을 충분한 메모리가 있다고 해도 계산 시간이 문제가 된다.

본 연구에서는, 웹 문서와 색인어를 이용하여 만든 행렬 m_{ij}은 대부분의 요소가 0인 희소행렬임을 이용하여 행렬의 요소가 0이 아닌 것끼리만 계산하여 메모리의 사용과 계산량을 큰 폭으로 줄여서 색인어간 연관도를 계산한다.

우선 데이터베이스를 1회 스캔하여 각 색인어마다 파일을 만든다. 이 파일에는 해당 색인어가 어떤 문서에 몇 번 출현하는지 정보가 입력된다. 이 파일들은 데이터베이스를 1회만 스캔하면 되므로 적은 시간 안에 모두 생성할 수 있다. 이 파일을 이용하여 특정 색인어가 출현하는 문서만 스캔하여 연관도를 계산할 수 있다. 예를 들어 색인어 "행주산성"에 대한 정보가 기록되어 있는 파일이 <표1>와 같다면, 112번, 1019번, 2080번과 같이 "행주산성"이 출현하는 문서만 읽어와서 그 문서에 출현하는 다른 색인어의 출현 횟수와 곱을 하여 연관도를 계산한다. <표2>에서 보는 것과 같이 행렬의 곱을 이용할 때에는 각 색인어에 대해 문서 수만큼 읽어와야 하므로 문서의 수가 많아질수록 계산시간은 급증하게 된다. 하지만 희소행렬임을 이용해 연관도를 계산할 때는 필요한 문서만 읽으면 되므로 계산시간이 상대적으로 많이 줄어들게 된다.

<표 1 "행주산성"이 출현하는 문서 번호와 출현 횟수 정보>

문서 번호	출현 횟수	문서 번호	출현 횟수
112	2	4568	5
1019	3	12354	12
2080	1	31586	3

<표2 연관도를 계산하는 소요 시간>

연관도 계산 방법	고양시 관련	경기도 관련
문서 수	6,825 개	119,459 개
행렬의 곱	229 초	529,000 초
희소행렬	8 초	4,320 초

5. 결론 및 향후과제

본 연구에서는 특정지역과 관련된 색인어간의 연관도 분석을 위해, 연관규칙을 이용한 방법과 연관 클러스터를 이용한 방법을 적용하였으며, "Kyonggi21Search"시스템에서는 연관 클러스터를 이용한 결과가 더 좋다는 것을 보였다. 또한 웹 문서와 색인어간의 행렬은 희소행렬임을 이용하여 연관 클러스터를 이용할 때 계산량을 큰 폭으로 줄이는 방법을 제시했다.

향후과제로는 MBR을 이용한 지도검색으로 웹 문서를 사용자에게 제공해 주는 인터페이스를 연구하고, 본 시스템에 좀 더 적합한 색인어간 연관도 분석 방법 등이 있다.

* 본 연구는 한국항공대학교 전자·정보통신·컴퓨터공학부 논리회로연구실(권용진 교수)과 일본 교토대학 대학원 사회정보학전공 Kambayashi연구실과의 국제공동연구의 일환이다.

참고문헌

[1] Baeza-Yates, Ribeiro-Neto "Modern Information Retrieval", Addison-wesley, 1999.
[2] GIS.com, http://www.gis.com
[3] 장정훈, 이룡, Y. Kambayashi, 권용진, "웹 정보 기반의 지역정보 시스템 구현 : "Kyonggi21 Search" 한국통신학회 추계학술발표논문집 vol 26, pp.252, 2002.
[4] 강현철, 한상태, 최종후, 김은석, 김미경, "SAS Enterprise Miner를 이용한 데이터마이닝-방법론 및 활용"1999년 1판
[5] http://ee.snu.ac.kr/~shim/
[6] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Dataminig: A performance perspective", IEEE Transactions on Knowledge and Data Engineering, 5(6), 12. 1993.