연관 웹 페이지 검색을 위한 e-아크 랭킹 메저
(e-Cohesive Keyword based Arc Ranking Measure for Web Navigation)

이 우 기 † 이 병 수 ††
(Wookey Lee) (Byoungsu Lee)

요 약 웹은 사용자에게 제품이나 정보를 제공할 수 있는 가장 커다란 매체로 성장하였으며, 또한 사용자에게는 필요 이상의 정보를 얻게 해주고 있다. 웹은 다량의 관련 정보들을 여러 웹 페이지들을 통해 표현하고 있으며, 현재 검색엔진들은 키워드들에 관련된 단일 페이지들만을 리스트화하여 보여주고 있다. 근본적으로 이러한 방법들로는 관련된 정보를 가지고 있는 페이지들의 쌍 및 연관된 웹 페이지들의 집합을 구조화하여 제공할 수 없다. 웹은 하나의 웹 페이지에 모든 관련 정보를 담는 범위를 넘어 관련된 정보 페이지들을 하이퍼링크로 서로 연결한 일련의 정보로 인식되고 있다. 따라서 본 논문에서는 새로운 링크 가중치 기반 검색 기법으로서 e-아크 메저에 관하여 제안하고자 하며, 이는 사용자가 입력한 키워드들과 관련된 페이지의 집합을 웹 사이트 안에서 찾아내는 연관 검색에 효과적이라는 것을 보이고, 실험을 통해 기존의 메저들 보다 그 효과성을 우월하다는 점을 입증하였다.

키워드 : 검색엔진, 키워드 기반 랭킹, e-코헤시브 아크 랭킹 메저, 유사도

Abstract The World Wide Web has emerged as largest media which provides even a single user to market their products and publish desired information; on the other hand the user can access what kind of information abundantly enough as well. As a result web holds large amount of related information distributed over multiple web pages. The current search engines search for all the entered keywords in a single webpage and rank the resulting set of web pages as an answer to the user query. But this approach fails to retrieve the pair of web pages which contains more relevant information for users search. We introduce a new search paradigm which gives different weights to the query keywords according to their order of appearance. We propose a new arc weight measure that assigns more relevance to the pair of web pages with alternate keywords present so that the pair of web pages which contains related but distributed information can be presented to the user. Our measure proved to be effective on the similarity search in which the experimentation represented the e-arc ranking measure outperforming the conventional ones.

Key words : Search engine, Keyword-based Ranking, e-Cohesive Arc Measure, Similarity

· 이 논문은 2008 한국컴퓨터종합학술대회에서 '연관 웹 페이지 검색을 위한 코헤시브 아크 메저'의 제목으로 발표된 논문을 확장한 것임
· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업(IITA-2008-C1090-0801-0031)의 연구결과로 수행되었음
· 이 논문은 인하대학교의 지원에 의하여 연구되었습니다.

† 종신회원 : 인하대학교 산업공학과 교수
wookeylee@gmail.com
†† 학생회원 : 인하대학교 산업공학과
leebyoungsu@gmail.com
논문접수 : 2008년 8월 27일
심사완료 : 2008년 12월 4일

정보과학회논문지 : 데이타베이스 제36권 제1호(2009.2)

1. 서 론

웹은 현재 가장 큰 정보 매체의 하나로 성장함에 따라 그에 따른 효과 또한 놀라운 속도로 확산되고 있다. 웹 사용자들은 자유롭게 웹에 접속할 수 있을 뿐만 아니라 사용자 자신의 정보를 공개하고 배포하기 위한 수단도 되고 있다. 대부분의 사회 조직이나 단체들은 웹 공간에 그들의 웹 페이지를 만들어 사용자들에게 제공하고 있다. 웹은 정보의 홍수를 이루고 있지만 그 광대한 양에 비하여 그 검색방법은 검색엔진에게 받은 일련의 웹 페이지가 대부분이다. 사용자는 이러한 검색 결과를 가지고 정말 원하던 내용인지 확인하려면 웹 사이트를 하나하나 다시 탐색해야 하며 이 과정에서 사용자들의 요구에 부합되지 못하는 경우가 많다.

일반적으로 웹 사용자들은 웹의 광대한 양의 정보들 중 자신이 원하는 정보를 뽑아내기 위해 검색엔진을 사용한다. 현재 많이 사용되고 있는 대부분의 검색엔진들은 키워드 기반 검색에 기초를 두고 있으며 이것은 사용자가 준 키워드 질의를 중심으로 각자 랭킹 기법을 통해 산출된 결과로 순위를 부여하여 하나의 리스트로 사용자에게 보여준다. 그러나 이러한 방법은 사용자가 준 키워드들을 가진 단일 페이지에만 초점을 맞추고 있으며 각 키워드들에 의해 검출된 페이지의 구조적 특징에는 신경 쓰지 않는다.

예를 들어 어떤 학생이 특정 분야의 연구활동을 하고 있는 대학 연구실에 관한 모집 정보를 필요로 한다고 하자 이때, 만약 그 학생의 관심분야가 "웹 구조화"라면 자신의 관심분야인 "웹 구조화"와 관련된 자신이 지원할 수 있는 대학을 찾아 보기 위해 검색창에 "대학교"와 "웹 구조화"라는 두 세 개 정도의 키워드를 검색창에 입력할 것이다. 이 경우 연구실 이름에 관련된 검색어가 모두 들어있지 않은 이상 현존하는 많은 검색엔진은 잘못된 결과를 보이기 쉽다. 왜냐하면 대학교 홈페이지에는 특정 연구 분야나 주제에 관한 정보를 가지고 있지 않으며, 또한 학과나 교수들 중에는 웹 구조화에 관련된 연구를 하고 있는 곳이 있다 하더라도 연구정보란에 소속 대학 이름을 명시하지 않는 경우에는 대학교 이름과 함께 단일 웹 페이지로 검색되지 않을 것이다.

또한 현존하는 대부분의 검색엔진은 기본적으로 키워드의 전후 관계보다 웹 페이지 안에 해당 키워드들의 존재에만 신경을 쓴다. 만약 좀더 상세한 내용의 키워드가 추가 된다면, 예를 들어 위의 상황에서 "대학교", "웹 구조화" 이외에 "미국"이나 "캐나다"와 같이 대학의 위치까지 추가한다면 정확도는 더 떨어질 것이다. 많은 학생들이 각자의 관심분야를 찾아 자신이 원하는 분야에 활발한 연구를 수행하는 교수님의 이름을 검색하여 그것을 바탕으로 다시 대학을 찾던지 아니면 각각의 대학교의 웹 사이트를 찾아 그 안에서 다시 관련 학과나 교수님을 찾아 보는 방법 등을 사용하여 관련 정보를 찾아볼 것이다. 그러나 이러한 방식의 방법은 매우 많은 시간을 소비해야 한다. 이러한 방식의 접근법 보다 더 좋은 대안은 관련 대학교 홈페이지와 교수나 연구실의 홈 페이지를 함께 제공하는 것이 좀 더 근원적인 문제 해결이 될 수 있는 점이다.

본 논문에서는 검색에 대한 새로운 패러다임을 제시하려 한다. 검색창에 입력되는 키워드들은 웹 검색자가 찾기 위한 정보에 관한 하나의 논리적 연관성을 가지고 있다고 보는 것이다. 즉, 이러한 질의는 계층적이거나 구조적 특징을 반영할 수 있다는 점에서 착안한 것이다. 이렇게 서로 논리적 연관성을 가지고 만들어진 키워드들을 본 논문에서는 코헤시브 질의(Cohesive query)라고 부르기로 한다.

본 논문에서는 검색결과가 순서화된 하나의 리스트에서 벗어나 웹 사용자가 필요로 하는 구조화된 웹 페이지들의 집합을 보여주는 프레임워크를 제안한다. 이는 사용자 질의의 연관성을 판단하기 위해 웹의 내용과 구조 이 두 가지 모두를 사용하며, 여기서 말하는 내용이란 웹 페이지들이 포함하는 키워드를 의미하며, 또한 구조란 같은 도메인이나 웹 사이트 안의 페이지들의 하이퍼링크로 연결된 일련의 결과를 의미한다. 웹의 각 페이지들의 구조는 논리적이며 개념적으로 서로 연관되어 있다고 가정한다. 그러므로 웹 검색자가 입력한 키워드 질의를 반영해주는 논리적 구조를 가지고 있는 웹 페이지의 쌍을 찾아 최소 서브그래프로 표현되는 구조적 검색결과를 잘 반영할 수 있도록 효과적인 아크 메저를 제시하는 것이 본 연구의 목표이다. 그러나 다른 검색 시스템과는 다르게 본 연구에서의 코헤시브 질의는 관련구조를 찾기 위해 두 개 이상의 키워드를 필요로 한다. 극단적으로는 최대 5개 이하의 키워드로서 관련 페이지를 설명할 수 있다는 실용적인 연구도 있다[1]. 물론 본 연구에서는 키워드의 숫자에 제한 받지 아니한다.

2. 관련 연구

하이퍼링크 구조를 통합된 키워드 기반 검색은 오래 전부터 연구되어 왔으며 다양한 검색엔진의 성공적인 사례들이 있다. 대표적으로 구글은 웹 페이지간에 링크 구조를 이용하여 각 웹 페이지를 점수화하는 페이지 랭크를 고안하여 구글 검색엔진에 사용하고 있다[2]. 또한 검색어 질의에 연관된 URL의 숫자를 줄이기 위해 웹 페이지 군집화를 적용한 제안도 있다[3]. 최근에는 이런 웹 페이지의 하이퍼링크 구조에서 높은 관련성을 가진 웹 페이지들을 찾기 위해 그래프 이론이 적용되고 있다[4,5]. 검색 시스템의 효과를 높이기 위해 사용자 질의의 결과로서 높은 관련성을 가진 웹 페이지들의 집합을 이용한 연구들이 발표되고 있으며, 이런 연구들의 대다수가 웹 사이트와 웹 페이지들의 하이퍼텍스트 링크구조가 중심이 되어 그 정보로 뽑아낸 중요 페이지들간의 최소 서브그래프나 서브트리를 사용자 질의의 결과로 제시한다[6,7].

본 연구를 가장 고무시킨 문헌은 주어진 임계값을 기준으로 모든 웹 페이지들의 쌍의 유사도 측정하는 방법[8]으로, 임계값을 넘는 유사 웹 페이지의 모든 쌍을 찾는 알고리즘이다. 그와는 달리 본 논문에서는 웹 페이지 안에 퍼져있는 키워드들의 관련 쌍을 찾는 기법을 제시하고자 한다.

또 다른 논문으로 웹 페이지들의 구성을 하나의 관련

원자결합인 정보 단위(information unit)의 개념으로 설명한 논문이 있다[9]. 이는 페이지 내의 질의 키워드의 유무에 기초한 키워드와 링크에 독립적인 인덱스를 주었으며 질의 가공 휴리스틱은 슈타이너 트리(Steiner Tree) 알고리즘을 사용하였다. 본 논문과 다른 점은 사용자가 입력한 질의 키워드를 모두 사용하여 최소 서브그래프에 사용하는 반면 본 논문의 시스템은 사용자가 입력한 첫 번째 질의 키워드를 통해 검색되어진 웹 사이트들을 기초로 나머지 키워드를 사용하여 웹 사이트의 최소 서브그래프를 구한다는 것이다.

또한 웹 사용자에게 제공하기 위한 데이타 검색 단위로서 연결 서브그래프의 유용성을 주장한 논문도 있다[10]. 본 논문에서는 키워드들의 가공을 위해 최소 서브그래프를 사용하며 또한 웹 페이지와 서브그래프 안의 키워드를 토대로 각각의 서브그래프 점수를 위해서도 사용한다. 그들은 같은 제작자에 의해 만들어진 페이지들간의 서브그래프 또한 고려하지만 앞에서 언급한 다양한 검색에 대한 적용을 위해 모든 질의 키워드에 같은 가중치 태그를 부여하는 반면 본 논문의 시스템은 질의 키워드에 다른 가중치를 부여하여 적용한다는 점이 크게 다르다.

또 다른 연구에서는 페이지 집합 순위법이라는 개념이 있다[11]. 이것은 각각의 검색된 웹 페이지들에 반하여 페이지들의 집합을 순위화하는 방법을 언급하였다. 그들은 연구에서 두 가지 특별한 도메인을 거론했는데 이것은 개요 질의와 상대적 질의이다. 그러나 그들의 프레임워크는 기본적으로 내용기반 분석을 사용한 페이지 집합 순위법 만을 사용하여 링크기반 분석 방법이 향후 연구 과제로 남아있다. 본 논문과 기본적으로 다른 점은 본 논문의 시스템은 질의 키워드의 순서에 따라 서로 다른 중요도를 부여한다는 점이고 또 다른 중요 요소는 본 논문의 시스템에서는 미리 결정된 URL의 집합에서 질의 키워드를 이용하여 구조화한다는 것이다.

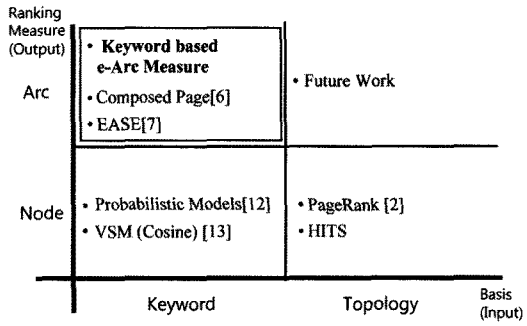


그림 1 연관된 기존연구와 Keyword based e-Arc Measure

그림 1은 연관된 다른 연구들과의 차이점을 정리하였다. 검색에서 키워드기반 검색과 구조기반 검색법이 있고, 한편 노드 가중치 부여법과 아크 가중치 부여법이 있는데 노드 가중치는 웹 페이지에 가중치를 부여하는 방법이며 아크 가중치는 하이퍼링크를 통해 페이지간의 관련성을 가중치화 하여 해당 아크에 값을 부여하는 방법이다. 본 연구는 키워드기반 아크가중치 부여 방법이다.

3. 시스템 아키텍처

본 연구에서는 키워드들의 입력 순서에 따른 중요도가 다르게 적용된다는 가정을 가지고 시작한다. 본 논문에서는 웹 검색자에게 입력된 첫 번째 키워드는 다른 키워드에 비하여 검색하고 싶은 정보에 관한 특별한 중요성을 내포하고 있다고 본다. 검색자들은 이런 중요 키워드를 기초로 더 상세한 다른 나머지 키워드를 적용하는 계층적 구조를 찾기 위해서 서로 다른 관심 분야에 따라 정보들이 계층적으로 조직화시킨 Goole Directory 나 Open Directory Project(ODP)와 같은 다양한 웹 디렉터리들을 사용하여 검색을 한다. 그러나 웹 검색자들은 자신이 원하는 특정 정보의 연결된 리스트 정보를 찾기 위해 이러한 디렉터리에서 제공한 웹 사이트의 호스트 URL을 찾아가 거기서부터 다시 특정 정보를 위한 검색을 하게 된다. 따라서 웹 검색자는 계층적 정보를 모으기 위해 반복적인 탐색노력을 계속 해야만 한다.

본 논문에서는 새로운 검색 패러다임을 제시한다. 이것은 첫 번째로 검색자의 관심과 관련된 웹 사이트의 리스트를 뽑아내고 각각의 웹 페이지들에 대해 검색자의 필요에 만족할 수 있는 웹 페이지의 쌍들을 찾는 웹 디렉터리적인 검색 결과를 보여준다는 것이다. 이러한 시스템은 다음의 두 가지 특징을 나타낸다:

가장 상위 개념이며 가장 중요한 첫 번째 키워드를 기초로 하여 웹 검색을 하여 관련 사이트들을 찾아낸다.

위에서 뽑아진 웹 사이트 각각에서 다시 세부적인 나머지 키워드를 적용하여 검색자가 필요로 하는 정보의 관련성을 가진 웹 페이지들을 찾아낸다.

본 논문에서 사용되는 기본 정의와 용어를 제시하고 그 적용대상 및 방법에 관하여 설명하겠다. 찾을 수 있는 모든 가능 키워드의 집합을 D라고 표시하고, 검색자의 질의 Q는 키워드들의 집합 $<k_1, k_2, ..., k_n>$을 나타내며, 앞에서 말한 중요 첫 번째 키워드는 k_1을 말한다. 또한 W는 키워드 k_1에 관련된 정보를 가지고 있는 웹 사이트 도메인의 집합을 의미 한다. 예를 들어 만약 k_1이 '회사'라는 첫 번째 키워드라면 W는 세계의 다양한 국내외의 유명 회사들의 홈페이지들을 포함하는 집합을 의미할 것이다.

각각의 웹 사이트의 URL $W_i \in W$에서 W_i의 도메인에 속해있는 노드들의 집합 즉, 웹 페이지들의 집합 V_i와 도메인 안의 웹 페이지들을 이어 주는 링크인 하이퍼링크들의 집합을 E_i라 할 때 유방향 그래프 G_i (V_i, E_i)라고 표현한다. V_i 집합은 같은 도메인 안의 웹 페이지만으로 제한한다. 왜냐하면 다른 도메인에서 결합된 웹 페이지들은 검색자를 잘못된 길로 의도할 우려가 있기 때문이다. 서로 다른 두 개의 대학의 웹 페이지가 서로 링크되어 있다면 협력관계의 대학일수도 있고 또 다른 많은 사회의 조직들과의 링크도 존재할 수 있다. 예를 들어 대학의 입학처와 다른 대학의 학과 소개와의 링크는 매우 잘못된 결과를 도출할 수 있다.

다음으로 각각의 W_i가 가지는 최소 서브그래프를 $V_i^m \subset V_i$이고 $E_i^m \subset Ei$일 때 $G_i^m(V_i^m, E_i^m)$라고 정의한다. V_i^m에서 W_i의 도메인 밖에 있는 페이지들의 집합은 제외되며 물론 W_i와 다른 도메인과의 링크 또한 제거된다. 이런 링크를 크로스 레퍼런스 링크라 한다.

키워드의 집합 D 안의 모든 키워드들은 하나의 최소 서브그래프와 매칭된다. 본 논문에서 다른 도메인과의 크로스 레퍼런스 링크와 도메인 안의 사이클 제거는 사용자의 질의와 독립적으로 실행된다. 따라서 이 작업은 질의 응답 시간을 줄이기 위해 선 처리 되어야 한다. 왜냐하면 검색자가 입력한 키워드 k_1의 결과로 광대한 양의 웹 사이트 URL들이 존재할 것이고 그 웹 사이트들의 최소 서브 그래프를 구하는 일은 많은 시간을 요구할 것이기 때문이다. 즉, 이것은 사용자 질의응답 시간을 늦추는 결과가 된다.

본 논문에서는 최소 서브그래프의 노드집합 V_i^m 안에 모든 웹 페이지들의 유사도를 측정하기 위해 새로운 기법을 개발하였고 이것을 입실론-코헤시브 아크 메저(Epsilon-Cohesive Arc Measure)라고 하며 이하 이-코헤시브 메저(E-Cohesive Measure)라 부르기로 한다.

3.1 대상 웹사이트 결정

첫 번째 키워드는 필요한 검색 도메인인 기본 웹 사이트 집합을 얻기 위해 사용된다. 이것은 단순 웹 질의로 얻어질 수 있다. 예를 들어 "어린이와 책"이라고 질의 한다면 기본 웹 사이트의 집합 W는 어린이와 관련된 모든 웹 사이트를 찾을 것이다. 다른 예로 "windows xp service pack"이라는 질의를 한다면 기본 웹 사이트의 집합 W는 마이크로 소프트의 윈도우와 관련된 웹 사이트들이 대부분을 포함할 것이다.

이러한 많은 웹 사이트들은 하이퍼링크로 서로 복잡하게 연결되어 있기 때문에 이 시간을 줄이기 위해 모든 사이클과 백링크를 제거하여 최소 서브 그래프를 만드는 과정은 검색자가 입력하는 키워드와는 독립적인 작업임으로 우선 처리 되어야 한다. 검색자가 입력한 키워드 k_1은 W를 얻는데 사용되며 일반적으로 W의 집합을 얻어내기 위해 본 논문에서는 Google Directory와 Open Directory Project(ODP)와 함께 구글 검색엔진의 결과를 사용하였다.

따라서 본 논문에서는 최소 서브 그래프에서 관련 웹 페이지들의 쌍을 찾는 질의 응답시간만을 고려한다.

3.2 이-코헤시브 메저(E-Cohesive Measure)

이-코헤시브 메저는 최소 서브그래프 G_i^m 안의 V_i^m의 쌍을 이어주는 모든 링크의 가중치를 구하는데 사용된다. 이 가중치는 서로 링크되어 있는 웹 페이지들이 검색자가 원하는 정보와 어느 정도의 연관성을 가지고 있는지를 평가하게 되며 각각의 질의 키워드들이 링크되어있는 웹 페이지들의 쌍에 잘 퍼져있을 때 높은 가중치를 부여한다. 이 메저는 전통적인 검색 시스템의 내용 분석 접근을 기반으로 하며 다음의 두 가지 단계를 거친다:

표 1 용어설명

용어	설명
Q	$\{k_1, k_2 \cdots\cdots k_m\}$
$tf(k_i)$	term-frequency for keyword k_i
$idf(k_i)$	inverse-document frequency for keyword k_i
ck^j	actual contribution of k to j node, where $ck^j = tf(k) * idf(k)$ for $j = 1, 2, \cdots, n$.
$f(p)$	feature vector for node p
$R_{a,b}$	arc weight from node a to node b
$\otimes$	convolution of two feature vectors

- 1단계: 이-코헤시브 메저를 사용하여 최소 서브그래프 G_i^m의 모든 E_i^m의 가중치를 0에서 1사이의 값으로 구한다.
- 2 단계: 1단계에서 구해진 가중치를 가지고 G_i^m 안의 모든 웹 페이지들의 하이퍼링크로 연결되어있는 쌍의 순위를 열거한다. 이 단계는 실질적인 최소 서브그래프 안에서의 웹 페이지들의 쌍을 랭킹한 결과가 제시된다.

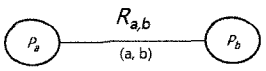


그림 2 웹 페이지 P_a와 P_b의 하이퍼링크(a, b) 쌍의 가중치 $R_{a,b}$

3.3 이-코헤시브 메저의 가중치 부여 방법

이-코헤시브 메저는 가중치를 구할 때 $TF*IDF$ (Term Frequency - Inverse Document Frequency)를 사용한다. 물론 키워드를 가중치 벡터로 표현하는 어떠한 방법이든 제한이 없이 본 연구의 방법에 적용될 수

있다. 또한 키워드 k_1은 이미 웹 사이트 W_i의 집합을 구하기 위해 사용되었기 때문에 이-코헤시브 아크 중요도를 구할 때는 제외되며 W_i 안의 노드 중 나머지 키워드 $<k_2, k_3, \cdots, k_n>$의 정보를 가지고 있는 노드를 추출하여 그 노드에 연결된 링크의 중요도를 구하게 된다.

이-코헤시브 메저를 구하기 위해 웹 페이지 $p_j \in V_i^m$ 안에 있는 k 키워드의 $TF*IDF$를 다음과 같이 구한다.

$$c_k^j = tf(k) * idf(k) \quad \text{for j = 1, 2, } \cdots \text{, n} \quad (1)$$

페이지 $p_j \in V_i^m$의 $f(p_j)$는 다음식 (2)에서와 같이 페이지 안의 각 키워드의 $TF*IDF$값의 벡터와 같으며 식 (3)의 $f(p_j)'$는 $f(p_j)$를 컴플리먼트(complement)한 것과 같다.

$$f(p_j) = (c_2^j, c_3^j, \cdots, c_n^j) \quad \text{for j = 1, 2, } \cdots \text{, n} \quad (2)$$

$$f(p_j)' = (1-c_2^j, 1-c_3^j, \cdots, 1-c_n^j) \text{ for j = 1, 2, } \cdots \text{, n} \quad (3)$$

$TF*IDF$ 값은 이런 텍스트 기반 데이타 분석의 수단으로 사용되는 방법으로 키워드의 숫자의 빈도를 이용하여 키워드와 페이지의 유사도를 반영하는 것으로 본다. 이-코헤시브 메저에서는 이 유사도를 이용하여 이번에는 링크로 연결되어 있는 페이지간의 유사도 가중치를 다시 산출하는 방법으로 이용한다. 이 가중치가 적은 페이지를 다음 과정인 이-코헤시브 메저 연산에서 제외시킴으로써 계산 양을 줄일 수 있다. 예를 들어 한 페이지의 중요도가 (0, 0, ⋯, 0)라면 키워드를 하나도 가지고 있지 않다는 것이고 이것은 명확히 이 페이지가 다음 과정에 필요하지 않다는 것을 의미한다.

코사인 유사도 메저[13]는 가장 널리 사용되는 유사도 측정기법 중에 하나이다. 이는 웹 페이지간의 유사도를 측정할 때 효과적인 결과를 보이지만 각 키워드가 퍼져 있는 웹 페이지들의 쌍, 즉 두 개의 페이지가 링크를 통해 연결되어 존재할 때 각각의 페이지에 키워드 또한 각각 다른 키워드를 가지고 있는 서로 연관성이 있는 쌍을 검색하는 것에는 좋지 않다. 이러한 쌍들은 서로 다른 페이지에 각각의 키워드들이 각각 존재하고 링크를 통해 연결돼있어 상호 보완적인 구조를 가진다. 따라서 이러한 구조에 더 높은 가중치를 부여하기 위해 이-코헤시브 메저를 사용하며, 그 식은 다음과 같다. a와 b는 그림 2에서 의 웹 페이지 P_a와 P_b를 의미한다.

$$R_{a,b} = f(a) \otimes f(b) \quad (4)$$

위의 곱 ⊗(convolution)은 다음과 같이 정의 된다.

$$R_{a,b} = f(a) \otimes f(b)$$

$$= (c_2^a, c_3^a, \cdots, c_n^a) \otimes (c_2^b, c_3^b, \cdots, c_n^b)$$

$$= \frac{f(a)\cdot f(b)'}{2|f(a)||f(b)'|} + \frac{f(a)'\cdot f(b)}{2|f(a)'||f(b)|}$$

$$= \frac{\sum_{x=2}^{n} c_x^a \times (1-c_x^b)}{2\sqrt{\sum_{x=2}^{n}(c_x^a)^2} \times \sqrt{\sum_{x=2}^{n}(1-c_x^b)^2}} + \frac{\sum_{x=2}^{n}(1-c_x^a)\times c_x^b}{2\sqrt{\sum_{x=2}^{n}(1-c_x^a)^2} \times \sqrt{\sum_{x=2}^{n}(c_x^b)^2}} \quad (5)$$

따라서 최종적인 알고리즘의 수행과정은 다음 그림 3의 알고리즘 의사코드와 같이 표현 할 수 있다.

```
1: Parse user query Q as {k1, k2, ... , kn }
2: Obtain the website URLs, W from k1
3: For each Wi ∈ W do
4: Obtain minimal acyclic subgraph Gi^m:(Vi^m, Ei^m)
5: for every web page pa ∈ Vi^m do
6:         construct f(pa)
7: for every pair pa, pb ∈ Ei^m do
8:         Ra,b = f(pa) ⊗ f(pb)
9: Present the top pair from each Wi to the user.
```

그림 3 이-코헤시브 메저 알고리즘

이상의 과정에 대해 세 개의 키워드로 검색하는 경우를 예를 들어 설명하겠다. 키워드 $Q = <k_1, k_2, k_3>$에서 중요키워드인 첫 번째 키워드 k_1를 사용하여 매칭되는 관련 도메인 집합 $W_i \in W$를 구한다. 이것으로 W_i의 모든 웹 사이트는 일단 k_1과의 유사성을 가지며 이렇게 도메인의 각 웹사이트들의 최소 서브그래프 $G_i^m(V_i^m, E_i^m)$를 구한다. 이제 각 서브그래프에서 남은 두 개의 k_2와 k_3의 $TF*IDF$값을 산출하여 각각의 도메인 안의 페이지 $p_j \in V_i^m$와 키워드간의 연관성을 찾아낸 후 이것을 바탕으로 두 페이지 $P_i, P_j \in E_i^m$의 연관성을 찾아낸다.

한 서브그래프 안의 웹 페이지 p_a, p_b, p_c 각각의 k_2와 k_3에 대한 $TF*IDF$값 $f(a) = (1, 0)$, $f(b) = (0, 1)$, $f(c) = (1, 1)$을 가지고 있다고 할 경우, 그 조합에 대한 결과는 표 1과 같다.

표 1 웹 페이지 p_a, p_b, p_c의 이-코헤시브 메저 결과 값

$R_{a,b} = f(a) \otimes f(b)$	$R_{a,b} = (1, 0) \otimes (0, 1) = 2$
$R_{b,a} = f(b) \otimes f(a)$	$R_{b,a} = (0, 1) \otimes (1, 0) = 2$
$R_{a,c} = f(a) \otimes f(c)$	$R_{a,c} = (1, 0) \otimes (1, 1) = 0.7$
$R_{c,a} = f(c) \otimes f(a)$	$R_{c,a} = (1, 1) \otimes (1, 0) = 0.7$
$R_{b,c} = f(b) \otimes f(c)$	$R_{b,c} = (0, 1) \otimes (1, 1) = 0.7$
$R_{c,b} = f(c) \otimes f(b)$	$R_{c,b} = (1, 1) \otimes (0, 1) = 0.7$
$R_{a,a} = f(a) \otimes f(a)$	$R_{a,a} = (1, 0) \otimes (1, 0) = 0$
$R_{b,b} = f(b) \otimes f(b)$	$R_{b,b} = (0, 1) \otimes (0, 1) = 0$
$R_{c,c} = f(c) \otimes f(c)$	$R_{c,c} = (1, 1) \otimes (1, 1) = 0$

표 2에서 보는 바와 같이 $TF*IDF$ 값이 (1, 0), (0, 1)이나 (0, 1), (1, 0)과 같은 웹 페이지의 쌍에 코사인 메저와는 다른 높은 가중치를 부여한다. 하지만 (1, 1), (1, 1)과 값은 좋은 쌍에는 나쁜 결과값이 나오는 것을 확인 할 수 있다. 이것은 $TF*IDF$ 값을 극단적인 0이나 1로 주어졌을 때 나오는 현상으로 이런 극단적인 현상을 막기 위해 $TF*IDF$ 값에 아주 작은 입실론(ε)값

표 2 이-코헤시브 메저와 코사인 메저의 Top-10 결과

Rank	*TF*IDF*		E-Cohesive	Cosine
1	(0, 0.7231)	(0.7735, 0)	0.97122	0
2	(0, 0.6673)	(0.8439, 0)	0.970025	0
3	(0, 0.9877)	(0.397, 0)	0.928936	0
4	(0, 0.3567)	(0.8439, 0)	0.915836	0
5	(0.1866, 0.5002)	(0.3395, 0)	0.914395	0.349521
6	(0, 0.7041)	(0.397, 0)	0.909308	0
7	(0, 0.7769)	(0.2927, 0)	0.897584	0
8	(0.3052, 0)	(0, 0.7041)	0.89158	0
9	(0.2927, 0)	(0, 0.6673)	0.884163	0
10	(0, 0.6024)	(0.3395, 0)	0.883455	0

을 더하거나 빼주어 가중치를 재 계산해 주며, 그 식은 다음과 같다.

$$\widehat{R_{a,b}} = \frac{\sum_{x=2}^{n} c_x^a \times (1 - c_x^b - \epsilon)}{2\sqrt{\sum_{x=2}^{n} (c_x^a)^2} \times \sqrt{\sum_{x=2}^{n} (1 - c_x^b - \epsilon)^2}} + \frac{\sum_{x=2}^{n} (1 - c_x^a - \epsilon) \times c_x^b}{2\sqrt{\sum_{x=2}^{n} (1 - c_x^a - \epsilon)^2} \times \sqrt{\sum_{x=2}^{n} (c_x^b)^2}} \quad (6)$$

그림 4는 입실론(ε)값의 변화에 따른 데이타의 변화이다. 이 결과를 봤을 때 입실론(ε)은 0.01 에서 0.05사이의 값이 적합하며 다음 4장 실험에서는 0.01의 값을 적용하여 실험을 하였다. 이렇게 모든 도메인 안에서의 키워드와 연관 페이지의 쌍을 찾아내어 각 도메인당 가중치가 높은 쌍을 검색결과로 제시한다.

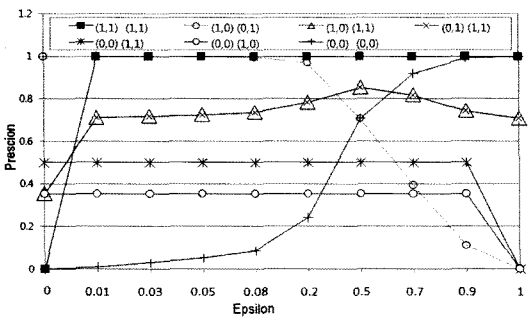


그림 4 입실론을 적용한 이-코헤시브 메저 값의 변화

4. 실험 및 분석 결과

본 논문에서는 이-코헤시브 메저를 효과적인 실험을 위해 두 가지 실험을 수행하였다. 즉, 가상의 웹 그래프를 대상으로 한 실험과 실제 웹 사이트에 대한 실험이 그것이다. 우선 가상 웹 그래프는 임의로 구해진 링크 464개를 가진 100개의 가상의 웹 페이지에서 실험 되었으며, 각 노드에 들어있는 키워드에 대한 *TF*IDF*값 또한 0에서 1사이의 값으로 임의로 주어졌다. 가상의 웹 그래프에서도 관련 키워드의 빈도에 따른 관련성이 높은 페이지들의 쌍이 높은 가중치 결과가 나왔으며, (1, 0), (0, 1)이나 (0, 1), (1, 0)과 같은 웹 페이지의 쌍에 코사인 메저와는 다른 높은 가중치를 부여한다.

처음 실험에서는 알고리즘의 정확한 결과치를 보기 위해 *TF*IDF* 값을 1과 0의 모든 조합으로 대입하여 그림 5와 같이 완전 그래프를 통해 그 결과치를 산출하였다.

그림 5에서 보는 바와 같이 키워드 2개를 모두 가지고 있는 쌍과 각각의 키워드를 하나씩 가지고 있는 쌍이 가장 높은 가중치를 가지며 동일한 하나의 키워드를 가지고 있는 쌍이 가장 낮은 가중치를 가진다는 것을 볼 수 있다.

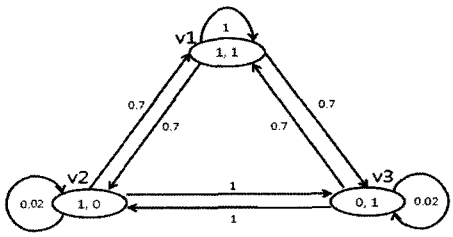


그림 5 완전 그래프에서의 링크 유사도

다음으로 *TF*IDF* 값을 임으로 0과 1사이 값을 부여하여 적용한 코사인 메저와 이-코헤시브 메저의 실험결과를 비교 분석해 보았다. 표 3은 이-코헤시브 메저의 상위 10개의 가중치에 따른 코사인 메저의 가중치이다.

표 3에서 이-코헤시브 메저에서 가장 높은 가중치를 가지고 있는 (33, 48) 링크의 *TF*IDF* 벡터는 각각 (0, 0.7231)와 (0.7735, 0)이다. 질의 키워드가 각각 두 개의 쌍 페이지에 하나씩 존재함으로써 서로 보완적인 페이지의 쌍을 이루고 있다. 이것은 앞에서 예를 들었던 대학의 연구실을 찾는 문제에 본 방식을 대입하여 설명하자면 33번 페이지는 대학의 홈 페이지를 의미하고 48번은 그 대학의 웹 구조화 연구에 관련된 페이지가 될 것이다. 이와 대조적으로 코사인 메저의 경우 이-코헤시브 메저의 결과값과는 전혀 다르게 거의 모든 가중치가 낮게 부여되었다. 이것은 이-코헤시브 메저와 코사인 메저는 전혀 다른 메저이며 각 페이지에 키워드가 하나씩 존재하는 서로 보완적인 페이지를 찾아주지 못한다는 것을 의미한다. 같은 첫 번째 키워드를 가지고 있음으로 두 페이지가 첫 번째 키워드에 관련된 높은 유사성을 가진다는 것을 의미할 뿐 구조적 의미를 찾아보기 힘들다.

다음으로 실제 웹 사이트에 대해 실험을 수행하였다. 이것은 웹 사이트는 질의: "conference information technology"라는 질의를 구글에 넣었을 때 얻어진 검색결과 10개를 선정하여 분석하였다. 이는 첫 번째 키워드

표 3 구글 웹 검색결과 및 웹 사이트 상세사항

#	Returned URL	Webpage text	Node	link
1	http://2009.informingscience.org	2009 Informing Science + Information Technology Education Joint Conference.	19	33
2	http://libra.msra.cn/ConferenceDetail.aspx	Conference On Information Technology Education	421	2791
3	http://site.aace.org/conf/	SITE, the Society for Information Technology & Teacher Education	31	422
4	http://www.secitc.eu/secitc2008/index.jspx	International Conference on Security for Information Technology and Communication	16	197
5	http://www.eit2006.org/	The 2006 IEEE International Electro/Information Technology Conference	25	314
6	http://cib.bau.tu-dresden.de/w78/	ITC@EDU about | submissions | program | venue | location | gallery.	104	327
7	http://www.gitma.org/	The Tenth Annual Global Information Technology Management Association	50	201
8	http://www.iccitbd.net/	Information about the ... Khulna University of Engineering & Technology	31	274
9	http://caise09.thenetworkinstitute.eu/	The 21st International Conference on Advanced Information Systems	13	144
10	http://wire.cs.nthu.edu.tw/itre2005/	3rd International Conference Sponsored by National Tsing Hua University Technically co-sponsored by IEEE ...	27	281

k_1을 의미하며 도메인을 선정하는데 사용되었다. 이후 키워드 k_2: "IEEE", k_3: "Chair"를 적용하였다.

$$Precision\ (a,\ b) = argmax_{(a,\ b)} R_{a,\ b} \quad (7)$$

본문에서는 도메인 중에서 두 웹 사이트(표 4의 8번째 및 10번째 웹 사이트)를 선정하여 이-코헤시브 메저(Epsilon Cohesive)와 코사인 메저(Cosine)[2,6] 그리고 글로벌 메저인 페이지 랭크(ArcPageRank)를 식 (7)과 같은 정확도(Precision)로 산출하여 그 효과성을 비교 분석하였다.

그림 6, 7은 각 웹사이트 내에서 이-코헤시브 메저의 상위 10개의 링크에 대한 코사인 메저와 페이지 랭크 결과값을 비교한 그래프로 시뮬레이션 결과와 유사한 결과를 보이는 것을 알 수 있다. 페이지 랭크 방법의 경우 질의와는 무관하게 인링크수에 따라 계산되므로 다소 랜덤하게 제시되는 것은 불가피하며 코사인의 경우 비슷하거나 상반되는 결과치를 보여주는데 이것은 위에서 설명한 바와 같이 키워드 값이 (1, 0), (0, 1) 쌍에서는 서로의 정확도(Precision)가 서로 상반되게 측정되며 나머지 쌍에서는 비슷한 결과를 보여주기 때문이다. 앞의 두 메저와는 대조적으로 본 방법은 정확도(Precision) 값들이 매우 우수한 결과를 보였으며, 노드 및 아크 수의 변화에 대해 일관성 있게 우수한 결과를 보이고 있다. 이러한 결과들은 본 연구방식에서 제안하는 구조화가 매우 효과적이고 유용하다는 점을 입증한 것이라고 볼 수 있다.

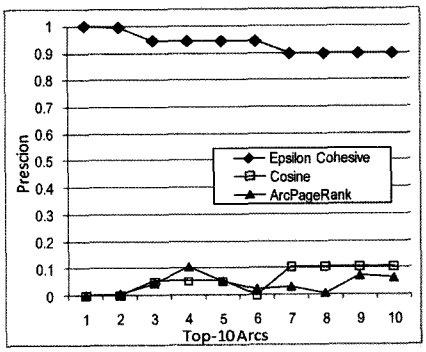


그림 6 wire.cs.nthu.edu.tw의 경우

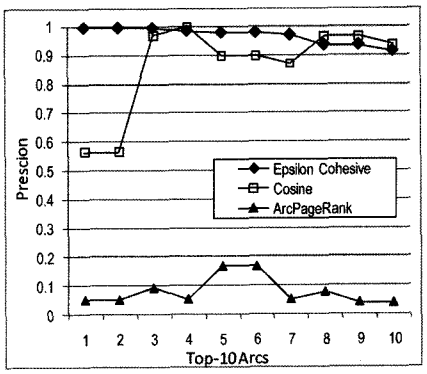


그림 7 www.iccitbd.net의 경우

5. 결 론

본 논문은 웹 검색의 결과로 현재의 검색엔진들이 보여주는 순위 리스트가 아닌 연관되어 있는 웹 페이지들의 쌍을 보여주는 새로운 메저를 개발하였다. 이 검색 기법에는 검색자가 입력한 키워드들 중 첫 번째 키워드에 많은 의미를 부여하여 검색의 범위를 한정 하였으며 이-코헤시브 메저를 사용하여 각 하이퍼링크들의 가중치를 구하였다. 이 메저는 각 웹 페이지가 하이퍼링크로

연결된 쌍들 중 키워드의 빈도수와 입력된 순서에 따라 검색자가 원하는 정보의 웹 서브 그래프를 찾아줄 수 있다는 것이다. 본 연구에서는 시뮬레이션과 실제 웹 페이지들에 대한 실험을 통해도 효과성을 입증하였다.

참 고 문 헌

[1] T. Phelps and R. Wilensky, "Robust Hyperlinks: Cheap, Everywhere, Now," Digital Documents and Electronic Publishing, *DDEP/PODDP*, pp. 28-43, 2000.

[2] L. Page and S. Brin, "The Anatomy of a Large Scale Hyper textual Web Search Engine," *WWW*, pp. 107-117, 1998.

[3] D. Gibson, J. M. Kleinberg and P. Raghavan, "Inferring Web Communities from Link Topology," *Hypertext*, pp. 225-234, 1998.

[4] R. Andersen, F. Chung, K. Lang, "Local Graph Partitioning using Page Rank Vectors," *FOCS*, pp. 475-486, 2006.

[5] S. Chakrabarti, A. Frieze and J. Vera, "The Influence of Search Engines on Preferential Attachment," *SODA*, pp. 293-300, 2005.

[6] R. Varadarajan, V. Hristidis, and T. Li, "Beyond Single-Page Web Search Results," *IEEE TKDE*, pp. 411-424, 2008.

[7] G. Li, B. C. Ooi, J. Feng, J. Wang, L. Zhou, "EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data," *SIGMOD*, pp. 903-914, 2008.

[8] R. J. Bayardo, Y. Ma and R. Srikant, "Scaling up All Pairs Similarity Search," *WWW*, pp. 131-140, 2007.

[9] W. Li, K. Candan, Q. Vu and D. Agrawal, "Retrieving and Organizing Web Pages by Information Unit," *WWW*, pp. 230-244, 2001.

[10] W. Lee and S. Lim, "Maximum Rooted Spanning Trees for the Web," *OTM*, pp. 1873-1882, 2006.

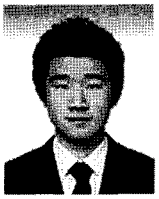
[11] T. Yumoto and K. Tanaka, "Page Sets As Web Search Answers," *ICADL*, pp. 244-253, 2006.

[12] M. Hammami, Y. Chahir and L. Chen, "WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis," *IEEE TKDE*, pp. 272-284, 2006.

[13] B. Neto and R. Baeza-Yates, *Modern Information Retrieval*, Addison-Wesley, 2001.

이 우 기

인하대학교 산업공학부 교수(현재). 서울대학교 산업공학과 학사, 석사 및 박사 UBC 방문교수. 한국정보과학회 및 경영정보학회 이사. Journal of IT&A 편집위원장(현재). 관심분야는 DB+IR, Web Mining, DW, EA 등



이 병 수

2007년 성결대학교 컴퓨터공학과 학사 현재 인하대학교 산업공학과 석사과정 관심분야는 정보검색, 웹마이닝, EA 등