태그를 이용한 웹 페이지간의 유사도 측정 방법
(Measuring Web Page Similarity using Tags)

강 상 욱 † 이 기 용 †† 김 현 규 ††† 김 명 호 ††††
(Sang Wook Kang) (Ki Yong Lee) (Hyeon Gyu Kim) (Myoung Ho Kim)

요 약 소셜 북마킹(social bookmarking)은 현재 웹에서 가장 활발한 트렌드 중의 하나이다. 소셜 북마크 시스템을 통해 사용자들은 원하는 웹 페이지에 그의 주제 또는 내용을 나타내는 태그(tag)들을 부착할 수 있다. 지금까지의 연구들은 주로 이러한 정보를 웹 검색을 향상시키는 데 사용해왔다. 본 논문에서는 웹 페이지에 부착된 태그들을 사용하여 두 웹 페이지 간의 의미적 유사도를 측정하는 방법을 제안한다. 웹 페이지는 다양한 종류의 멀티미디어 데이터로 구성되어 있기 때문에, 웹 페이지 내부에 포함된 데이터를 사용하여 웹 페이지 간의 유사도를 측정하는 것은 매우 어려운 일이다. 하지만 사용자들에 의해 웹 페이지에 부착된 태그들을 사용하면 웹 페이지 간의 유사도는 매우 효과적으로 측정될 수 있다. 본 논문에서는 WSET (Web Page Similarity Based on Entire Tags)라 하는, 태그에 기반하여 웹 페이지 간의 유사도를 측정하는 새로운 방법을 제안한다. 실험 결과는 제안하는 방법이 기존 방법에 비해 더 좋은 결과를 나타냄을 보였다.

키워드 : 웹 페이지 간 유사도, 태그, 소셜 북마크, WWW

Abstract Social bookmarking is one of the most interesting trends in the current web environment. In a social bookmarking system, users annotate a web page with tags, which describe the contents of the page. Numerous studies have been done using this information, mostly on enhancing the quality of web search. In this paper, we use this information to measure the semantic similarity between two web pages. Since web pages consist of various types of multimedia data, it is quite difficult to compare the semantics of two web pages by comparing the actual data contained in the pages. With the help of social bookmarks, this comparison can be performed very effectively. In this paper, we propose a new similarity measure between web pages, called Web Page Similarity Based on Entire Tags (WSET), based on social bookmarks. The experimental results show that the proposed measure yields more satisfactory results than the previous ones.

Key words : Web page similarity, Tag, Social Bookmarks, WWW

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2009-0083055)

† 정 회 원 : 삼성전자 무선사업부 사원
swkang@dbserver.kaist.ac.kr
†† 정 회 원 : KAIST 전산학과 연구조교수
kiyong.lee@gmail.com
(Corresponding author임)
††† 정 회 원 : KAIST 전산학과
hgkim@dbserver.kaist.ac.kr
†††† 종신회원 : KAIST 전산학과 교수
mhkim@dbserver.kaist.ac.kr

논문접수 : 2009년 7월 16일
심사완료 : 2010년 1월 19일

1. 서 론

웹 페이지는 텍스트, 그림, 동영상, 하이퍼링크 등 다양한 종류의 멀티미디어 데이터로 구성된다. 지금까지 이러한 웹 페이지 내부의 데이터를 사용하여 웹 검색의 질을 높이기 위한 연구들이 있어왔다. PageRank[1]와 Hyperlink-Induced Topic Search(HITS)[2]는 웹 페이지들의 하이퍼링크 정보를 사용하여 웹 페이지들의 검색 결과에 순위를 매기는 대표적인 예이며, [3]은 웹 사용자가 방문한 웹 페이지의 텍스트 데이터를 확률적 모델에 적용하여 개인화 검색을 향상시키는 예이다.

그러나 웹 페이지 내부의 정보를 이용하여 웹 검색의 질을 높이는 데에는 한계가 있다. 만약 어떤 웹 페이지가 이미지 파일이 더 중요한 정보를 포함하고 있다면, 앞서 언급한 텍스트 데이터 기반의 개인화 검색 모델은

정확한 검색 결과를 만들어내지 못할 것이다. 또한, 어떤 웹 페이지를 가리키고 있는 하이퍼링크가 많지 않다면, 해당 페이지가 아무리 사용자에게 중요한 웹 페이지라고 하더라도, [1]과 [2]에 의해서는 이 페이지가 검색 결과의 상위에 포함되기 어렵다. 이러한 문제점을 극복하기 위해 웹 페이지 외부의 정보를 이용하려는 연구들이 진행되어왔다. Chirita et al.[4]는 웹 페이지들에 대한 분류 정보를 공개적으로 제공하고 있는 Open Directory Project(ODP)를 사용하여 개인화 검색을 개선하려고 하였으며, Bao et al.[5]는 딜리셔스(delicious.com)[6]와 같은 소셜 북마킹 서비스로 다수의 웹 사용자들이 웹 페이지에 부착한 태그 정보를 사용하여 웹 검색을 개선하고자 하였다. 이렇게 웹 페이지 외부의 정보를 이용하는 연구들은 주로 효율적인 웹 검색 또는 개인화 검색의 개선에 목표를 두고 있다.

본 논문에서는 이러한 웹 페이지 외부의 정보를 웹 페이지 간의 의미적 유사도(semantic similarity)를 측정하는 데 사용한다. 웹 페이지 간의 의미적 유사도란 주어진 웹 페이지들이 얼마나 비슷한 주제 또는 내용을 다루고 있는지를 측정하는 척도이다. 만약 두 웹 페이지가 모두 '프로그래밍'과 관련된 주제를 다루고 있다면 두 페이지는 유사한 것이며, 만약 한 페이지는 '음식'과 관련된 주제를, 다른 페이지는 '프로그래밍'과 관련된 주제를 다루고 있다면 두 페이지는 유사하지 않은 것이다. 지금까지 웹 페이지 간의 유사도를 측정하는 연구는 많이 이루어지지 않았다. HITS[2]에서는 주어진 웹 페이지와 유사한 웹 페이지들을 검색하기 위해 웹 페이지들 간의 하이퍼링크 정보를 사용하였으며, SSR[5]는 유사한 웹 페이지들을 검색하기 위해 웹 페이지들에 부착된 태그들을 사용하였다.

웹 페이지 간의 유사도를 측정하기 위해 사용할 수 있는 방법에는 몇 가지가 있다. 먼저 웹 페이지가 가지고 있는 텍스트 데이터를 통해 웹 페이지 간의 유사도를 측정하는 방법이다. 이 방법은 IR 분야에서 두 문서 간의 유사도를 비교하기 위해 주로 사용되는 방법인 단어 벡터 모델(term vector model)을 사용한다. 이 방법은 비교하고자 하는 웹 페이지가 가지고 있는 텍스트 데이터로부터 단어 출현 빈도 벡터(term frequency vector)를 생성한다. 그 후, 단어 출현 빈도 벡터 간의 코사인 유사도(cosine similarity)를 계산함으로써 두 문서 간의 유사도를 구한다. 하지만 위 방법은 웹 페이지가 텍스트 정보 외에 다른 형태의 데이터를 포함하고 있는 경우에는 웹 페이지 간의 유사도를 바르게 측정하기 어렵다. 예를 들어, 반 고흐의 어떤 그림만을 가지고 있는 웹 페이지와, 이 그림에 대한 설명을 텍스트 데이터의 형태로 가지고 있는 웹 페이지간의 유사도를 구하는 경우를 생각해보자. 이 경우, 그림만을 가지고 있는 웹 페이지로부터는 단어 출현 벡터를 만들 수 없기 때문에 두 웹 페이지간의 유사도를 측정하기 어렵다. 이는 그림뿐만 아니라 텍스트가 아닌 다른 형태의 데이터에 대해서도 마찬가지이다.

이 외에 HITS[2]에서 제안한 유사 페이지 검색 방법을 생각해 볼 수 있다. HITS 알고리즘은 주어진 페이지가 있을 때, 이 페이지와 하이퍼링크로 연결된 페이지들을 모아서 하나의 그룹으로 만든다. 그리고 이 그룹 내에서 하이퍼링크 정보를 분석하여 가장 권위(authoritative)있다고 판단되는 페이지를 주어진 페이지와 가장 유사한 페이지라고 판단한다. 그러나 이 방법은 비교 대상이 되는 두 웹 페이지가 하이퍼링크로 연결되어 있지 않다면, 두 웹 페이지 간의 유사도를 측정할 수 없다. 그리고 아무리 다루는 주제가 유사한 페이지라 하더라도 두 페이지를 가리키고 있는 하이퍼링크의 수에 매우 차이가 크다면, 두 페이지는 유사한 페이지로 판단될 가능성이 매우 낮다.

이와 같이 웹 페이지 내의 정보를 이용하여 웹 페이지 간의 유사도를 측정하는 데에는 상당한 제약이 있다. 본 논문에서는 웹 페이지 내부의 정보가 아닌 웹 페이지 외부의 정보를 이용하여 웹 페이지 간의 유사도를 측정하는 방법을 제안한다. 본 논문에서 제안하는 방법은 비교의 대상이 되는 웹 페이지에 부착된 태그(tag)들을 사용하여 웹 페이지 간의 유사도를 측정한다. 태그란 주어진 웹 페이지가 어떤 주제나 내용을 다루고 있는가를 나타내는 하나의 단어 또는 키워드이다. 현재 많은 수의 웹 사용자들이 딜리셔스(delicious.com)와 같은 소셜 북마킹 시스템을 통해 웹 페이지에 그의 주제 또는 내용을 나타내는 태그들을 부착하고 있다. 예를 들어, 'http://www.naver.com'과 같이 국내에 잘 알려진 검색 사이트에 대해서는 '검색', '네이버'와 같은 태그가 부착될 수 있다. 이와 같이 웹 페이지에 부착된 태그 정보를 이용하면 기존의 방법보다 더 효과적으로 웹 페이지 간의 의미적 유사도를 측정할 수 있다. 본 논문에서는 태그를 사용하여 웹 페이지 간의 유사도를 측정하는 방법인 Web Page Similarity Based on Entire Tags (WSET)을 제안하고, 딜리셔스(delicious.com)로부터 얻은 데이터를 사용하여 수행한 실험 결과를 보인다. 실험 결과는 제안하는 방법이 기존의 방법보다 만족스러운 결과를 얻을 수 있음을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 태그를 사용한 기존의 유사도 측정 방법을 설명하고, 3장에서는 제안하는 방법을 자세하게 설명한다. 4장에서는 실험 결과를 분석하고, 5장에서는 본 연구의 결과를 정리하고 요약한다.

2. 기존 연구

1장에서 언급한 웹 페이지 내의 정보를 이용하여 웹 페이지 간의 유사도를 측정하는 방법 외에도, 본 논문과 유사하게 웹 페이지에 부착된 태그들을 사용하여 웹 페이지 간의 유사도를 측정하는 방법이 제안되었다. Social-SimRank(SSR)[5]는 웹 페이지에 부착된 태그들 간의 유사도를 측정하는 방법이다. 비록 SSR이 웹 페이지 간의 유사도를 측정하기 위하여 직접적으로 제안된 방법은 아니지만, SSR를 사용하면 태그 간의 유사도를 바탕으로 웹 페이지 간의 유사도를 구할 수 있다. SSR은 그의 결과로 태그 간의 유사도를 나타내는 2차원 행렬 S_A를 만들어낸다. N_A를 태그의 개수라 할 때 S_A는 $N_A \times N_A$의 크기를 가지며, 각 원소 $S_A(t_i, u_j)$는 두 태그 t_i와 u_j 간의 유사도를 나타낸다. 두 웹 페이지 P와 Q가 있고 각각의 태그를 $[t_1, t_2, ..., t_n]$와 $[u_1, u_2, ..., u_m]$라고 하자. SSR를 사용하여 P와 Q의 유사도 $S_p(P, Q)$를 구하기 위해서는 다음의 식과 같이 P와 Q가 가진 각각의 태그 간의 유사도를 모두 합하면 된다.

$$S_p(P_1, Q_1) = S_A(t_1, u_1) + S_A(t_1, u_2) + \cdots + S_A(t_n, u_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} S_A(t_i, u_j) \quad (1)$$

위와 같은 방법을 사용하면 웹 페이지 간의 유사도를 매우 쉽게 계산할 수 있다. 하지만 이 방법은 다음과 같은 문제점을 가지고 있다. 태그는 사람이 일상에서 사용하는 단어로서 여러 가지의 의미를 가질 수 있다. 예를 들어 'java'라는 태그는 프로그래밍 언어인 JAVA를 나타낼 수도 있고, 인도네이사의 섬 이름을 나타낼 수도 있다. 이와 같은 태그는 다른 어떤 태그와 같이 사용되었느냐에 따라 완전히 다른 의미를 가질 수 있다. 하지만 SSR을 사용하는 위의 방법은 어떤 태그가 다른 어떤 태그와 사용되었느냐에 따라 의미가 달라질 수도 있다는 것을 고려하지 않는다. 예를 들어, [java, programming]와 [java, travel]을 태그로 가진 두 웹 페이지 간의 유사도를 측정한다고 하자. 이 때, 전자에서의 'java'는 프로그래밍 언어인 JAVA로서 사용되었고 후자에서의 'java'는 인도네시아의 자바 섬 여행의 의미로 사용되었다고 가정하자. 따라서 두 웹 페이지는 서로 다른 주제를 다루고 있으며, 유사하지 않다고 판단되어야 한다. 그러나, SSR을 사용하면 두 웹 페이지 간의 유사도는 S_A(java, java) + S_A(java, travel) + S_A(programming, java) + S_A(programming, travel)와 같은 식으로 계산된다. 여기서 'java'와 'java' 또는 'java'와 'programming'은 서로 매우 유사한 태그라고 판단될 수 있기 때문에 S_A(java, java)나 S_A(programming, java)는 매우 높은 유사도 값을 가질 것이다. 반면 'programming'과 'travel'은 거의 연관되지 않는 태그이기 때문에 0에 가까운 유사도 값을 가질 것이다. 하지만 위의 덧셈을 계산하게 되면 최종적으로 0에 가까운 값이 아닌 높은 값이 나와 두 페이지가 유사하다고 판단될 가능성이 높다. 이것은 SSR을 사용하는 방법이 'java'가 'programming'과 같이 사용되었을 때와 'travel'과 같이 사용되었을 때, 각각 다른 의미를 가진다는 것을 고려하지 않기 때문이다. 이에 따라 SSR을 사용하는 방법은 다양한 의미를 가질 수 있는 태그가 포함될 경우 정확하지 못한 유사도를 나타낼 가능성이 높다.

3. Web page Similarity based on Entire Tags 방법

본 장에서는 본 논문이 해결하고자 하는 문제를 정의하고, 본 논문이 제안하는 웹 페이지 간의 유사도 측정 방법인 Web Page Similarity based on Entire Tags (WSET)에 대하여 자세히 설명한다.

3.1 문제 정의

두 웹 페이지 P와 Q가 주어져 있고, 각각에 부착된 태그를 $[t_1, t_2, ..., t_n]$와 $[u_1, u_2, ..., u_m]$라고 하자. 웹 페이지 P와 Q 간의 유사도를 $S_\sigma(P, Q)$라 할 때, 본 논문에서는 $[t_1, t_2, ..., t_n]$와 $[u_1, u_2, ..., u_m]$를 사용하여 $S_\sigma(P, Q)$를 정의한다.

3.2 Separable Mixture Model (SMM)

본 논문에서 제안하는 방법은 웹 페이지 간의 유사도를 구하기 위해 Separable Mixture Model(SMM)[7]을 이용한다. SMM은 동시 발생 데이터(co-occurrence data)를 위한 통계적 모델이다. 여기서 동시 발생 데이터란 동시에 발생하는 두 가지 다른 종류의 데이터를 뜻한다. 예를 들어, 문서와 그에 대한 키워드는 각각 문서의 집합과 키워드의 집합에서 발생한 데이터로 이루어진 동시 발생 데이터로 볼 수 있다. 웹 페이지와 태그의 경우에도 각각 웹 페이지의 집합과 태그의 집합에서 발생한 데이터로 이루어진 동시 발생 데이터로 볼 수 있다. 서로 다른 주제를 나타내는 K개의 추상 클래스가 주어졌다고 하자. 주어진 동시 발생 데이터의 집합에 대해 SMM을 구축하면, SMM은 주어진 동시 발생 데이터들을 각각 K개의 추상 클래스로 분류한 뒤, 그의 결과로 (1) 각각의 추상 클래스가 발생할 확률과 (2) 동시 발생 데이터를 구성하는 두 가지 종류의 데이터 각각이 K개의 추상 클래스 각각에 대해 나타날 조건부 확률을 알려준다. 이로부터 어떤 임의의 동시 발생 데이터가 K개의 추상 클래스 각각에 대해 나타날 확률을 구할 수 있다.

이 때, 같은 추상 클래스에 대해 나타날 확률이 높은

데이터들은 같은 주제에 대한 것일 확률이 크며 유사하다고 볼 수 있고, 그렇지 않은 데이터들은 유사하지 않다고 볼 수 있다. 그리고 여러 추상 클래스에 걸쳐 나타날 확률이 높은 데이터들은 여러 주제에 관한 것이라고 볼 수 있다. 웹 페이지와 태그의 경우, 같은 추상 클래스에 대해 나타날 확률이 높은 웹 페이지 및 태그들은 유사한 주제에 대한 것일 확률이 높으며, 어떤 태그가 여러 추상 클래스에 걸쳐 나타날 확률이 높다면 해당 태그는 여러 주제에 관해 있을, 즉 여러 의미를 가지고 있을 가능성이 높다.

3.3 웹 페이지와 태그에 SMM 적용하기

앞 절에서 설명한 바와 같이 웹 페이지와 태그를 동시 발생 데이터로 보고 SMM을 적용하면, 구축된 SMM으로부터 각각의 웹 페이지와 태그가 K개의 추상 클래스 각각에 대해 나타날 조건부 확률을 구할 수 있다. K개의 추상 클래스가 C_1, C_2, ..., C_K로 미리 주어져 있다고 하자. P_1, P_2, ..., P_u의 웹 페이지와 t_1, t_2, ..., t_v의 태그로 이루어진 주어진 동시 발생 데이터에 SMM을 적용하면, 구축된 SMM으로부터 다음의 값을 얻을 수 있다.

- $p(C_\alpha)$: 클래스 α가 발생할 확률 ($1 \le \alpha \le K$).
- $p(P_i|C_\alpha)$: 클래스 α가 발생했을 때, 웹 페이지 P_i가 나타날 조건부 확률 ($1 \le i \le u$).
- $p(t_i|C_\alpha)$: 클래스 α가 발생했을 때, 태그 t_i가 나타날 조건부 확률 ($1 \le i \le v$).

그림 1은 주어진 웹 페이지와 태그의 집합에 SMM을 적용한 간단한 예이다. 그림 1을 보면 4개의 웹 페이지 A, B, C, D와 3개의 태그 'programming', 'java', 'tour'가 존재한다. 각 페이지는 두 개의 태그를 가지고 있는데, A, B, C는 각각 'programming'과 'java' 그리고 D는 'java'와 'tour'를 태그로 가지고 있다. 두 개의 추상 클래스인 클래스 1과 클래스 2가 주어졌을 때, 위의 데이터에 SMM을 적용하면 오른쪽 그림과 같은 결과를 얻게 된다. 클래스 1은 자바 프로그래밍에 관한 주제를 나타내며, 클래스 2는 자바 섬 여행에 관한 주제를 나타낸다. 그림 1의 오른쪽 그림에서 총 8개의 동시 발생 데이터 (A, 'programming'), (A, 'java'), (B, 'programming'), (B, 'java'), (C, 'programming'), (C, 'java'), (D, 'java'), (D, 'tour') 중 앞의 6개가 클래스 1로 분류되었으며, 나머지 2개는 클래스 2로 분류되었다.

표 1 그림 1의 결과 값

	Class 1	Class 2
$p(C_\alpha)$	0.75	0.25
$p('programming'|C_\alpha)$	0.5	0
$p('java'|C_\alpha)$	0.5	0.5
$p('tour'|C_\alpha)$	0	0.5
$p(A|C_\alpha)$	0.33	0
$p(B|C_\alpha)$	0.33	0
$p(C|C_\alpha)$	0.33	0
$p(D|C_\alpha)$	0	1.0

구축된 SMM으로부터 아래 표 1과 같은 값들을 얻게 된다. 표 1은 각 클래스가 발생할 확률, 각 클래스에 대해 4개의 웹 페이지 각각이 나타날 확률, 각 클래스에 대해 3개의 태그 각각이 나타날 확률을 나타낸다.

표 1을 보면, 그림 1의 오른쪽 그림이 나타내는 바와 같이 총 8개의 동시 발생 데이터 중 6개가 클래스 1로 분류되어 $P(C_1) = 0.75(= 6/8)$이 되었다. 이와 유사하게 $P(C_2) = 0.25(= 2/8)$이다. 그리고 클래스 1로 분류된 6개의 동시 발생 데이터에 대해 3개의 웹 페이지 A, B, C가 2번씩 나타났으므로 $P(A|C_1) = P(B|C_1) = P(C|C_1) = 1/3$이다. 또한 위의 예의 3개의 태그 중 'programming'과 'java'는 클래스 1에 속하는 6개의 동시 발생 데이터 중 각각 3번씩 나타났으므로 $P('programming'|C_1) = P('java'|C_1) = 1/2$이다. 이와 유사하게 'java'와 'tour'는 클래스 2에 속하는 2개의 동시 발생 데이터 중 각각 1번씩 나타나므로 $P('java'|C_2) = P('tour'|C_2) = 1/2$이다. 여기서, 'java'라는 태그는 다른 태그와는 달리 클래스 1과 클래스 2에 대해 모두 0보다 크게 나타날 확률 $P('java'|C_1) = 1/2$과 $P('java'|C_2) = 1/2$을 가짐에 유의해야 한다. 이것은 'java'가 클래스 1과 클래스 2 각각의 주제에 대해 두 가지 의미로 사용되었다는 것을 나타낸다. 다음 장에서 설명할 본 논문에서 제안하는 방

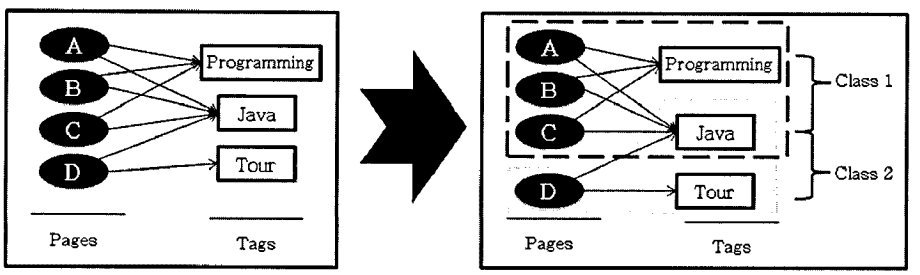

그림 1 웹 페이지와 태그를 SMM에 적용하는 간단한 예

법은 각 웹 페이지가 각 클래스에 대해 나타날 조건부 확률 $p(P_i|C_\alpha)$은 사용하지 않고, 각 클래스가 발생할 확률 $p(C_\alpha)$ 및 각 태그가 각 클래스에 대해 나타날 조건부 확률 $p(t_i|C_\alpha)$만을 사용한다.

3.4 Web page Similarity based on Entire Tags (WSET)

앞에서 설명한 $p(C_\alpha)$과 $p(t_i|C_\alpha)$을 사용하면, 두 개의 태그 t_1과 u_1가 같은 클래스에서 나타날 확률 $p(t_1 \wedge u_1)$을 다음과 같이 계산할 수 있다.

$$p(t_1 \wedge u_1) = p(t_1 \mid u_1)p(u_1) \quad (2)$$
$$\sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(C_\alpha \mid u_1) \cdot p(u_1)$$
$$\sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)\frac{p(u_1 \mid C_\alpha)p(C_\alpha)}{p(u_1)} \cdot p(u_1)$$
$$\sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(u_1 \mid C_\alpha) \cdot p(C_\alpha)$$

위와 같은 방법으로 $p(u_1 \wedge t_1)$를 계산하면 $p(t_1 \wedge u_1)$와 같다는 것을 알 수 있다. 따라서, 두 태그 t_1과 u_1가 같은 클래스에서 나타날 확률은 식 (3)과 같이 표현된다.

$$p(t_1 \wedge u_1) = p(u_1 \wedge t_1) = \sum_{\alpha=1}^{C} p(C_\alpha)p(t_1 \mid C_\alpha)p(u_1 \mid C_\alpha) \quad (3)$$

아래의 식 (4)는 식 (3)을 3개의 태그 t_1, t_2, u_1로 확장한 것이다. 세 개의 태그 t_1, t_2, u_1가 같은 클래스에서 나타날 확률 $p(t_1 \wedge t_2 \wedge u_1)$은 다음과 같이 계산할 수 있다.

$$p(t_1 \wedge t_2 \wedge u_1) = p(t_1 \wedge t_2) \cdot p(u_1 \mid t_1 \wedge t_2)$$
$$= p(t_1) \cdot p(t_2 \mid t_1) \cdot p(u_1 \mid t_1 \wedge t_2)$$
$$= p(t_1) \cdot p(t_2 \mid t_1) \cdot \frac{p(t_1 \wedge t_2 \mid u_1)p(u_1)}{p(t_1 \wedge t_2)}$$
$$= \sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(t_2 \mid C_\alpha)p(C_\alpha) \cdot \frac{p(t_1 \wedge t_2 \mid C_\alpha)p(C_\alpha \mid u_1)p(u_1)}{p(t_1 \wedge t_2 \mid C_\alpha)p(C_\alpha)}$$
$$= \sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(t_2 \mid C_\alpha)p(C_\alpha) \cdot \frac{p(C_\alpha \mid u_1)p(u_1)}{p(C_\alpha)}$$
$$= \sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(t_2 \mid C_\alpha)p(C_\alpha) \cdot \frac{p(u_1 \mid C_\alpha)p(C_\alpha)}{p(u_1)} \cdot \frac{p(u_1)}{p(C_\alpha)}$$
$$= \sum_{\alpha=1}^{C} p(t_1 \mid C_\alpha)p(t_2 \mid C_\alpha)p(u_1 \mid C_\alpha)p(C_\alpha) \quad (4)$$

위의 식 (3)과 (4)로부터 (n + m)개의 태그 t_1, t_2, ..., t_n, u_1, u_2, ..., u_m가 같은 클래스에서 나타날 확률 $p(t_1 \wedge t_2 \wedge \ldots \wedge t_n \wedge u_1 \wedge u_2 \wedge \ldots \wedge u_m)$은 다음과 같이 계산할 수 있다.

$$p(t_1 \wedge t_2 \wedge \cdots \wedge t_n \wedge u_1 \wedge u_2 \wedge \cdots \wedge u_m)$$
$$= \sum_{\alpha=1}^{C} \prod_{i=1}^{n} p(t_i \mid C_\alpha) \cdot \prod_{j=1}^{m} p(u_j \mid C_\alpha) \cdot p(C_\alpha) \quad (5)$$

두 웹 페이지 P와 Q의 태그들을 각각 $[t_1, t_2, \ldots, t_n]$와 $[u_1, u_2, \ldots, u_m]$라고 할 때, P와 Q의 유사도 $S_\sigma(P,Q)$는 t_1, t_2, ..., t_n, u_1, u_2, ..., u_m가 같은 클래스에서 나타날 확률에 태그의 개수인 (n + m)근을 취해 기하 평균을 얻은 값이다. 따라서 웹 페이지 P와 Q의 유사도는 최종적으로 다음의 식 (6)과 같이 정의된다.

$$S_\sigma(P,Q) = \sqrt[n+m]{p(t_1 \wedge t_2 \wedge \cdots \wedge t_n \wedge u_1 \wedge u_2 \wedge \cdots \wedge u_m)} \quad (6)$$

그림 2는 제안하는 방법을 나타내는 슈도 코드(pseudo code)이다.

4. 실험 결과

4.1 샘플 데이터

```
function similarity(P, Q: 비교 대상이 되는 웹 페이지)
{
        /* SMM으로 얻어진 결과 */
        C1, C2, …, CK: 주어진 K개의 추상 클래스
        p(Cα): 클래스 Cα가 발생할 확률 (1 ≤ α ≤ K)
        p(ti|Cα): 클래스 Cα가 발생했을 때, 태그 ti가 나타날 조건부 확률 (1 ≤ α ≤ K)

        /* P와 Q의 유사도 계산 */
        t1, t2, ..., tn: P에 부착된 태그
        u1, u2, ..., um: Q에 추착된 태그
        value = Σ_{α=1}^{C} Π_{i=1}^{n} p(ti|Cα) · Π_{j=1}^{m} p(uj|Cα) · p(Cα)
        value = (n+m)√value

        return value
}
```

그림 2 제안하는 방법

제안하는 방법인 WSET의 성능을 측정하기 위해, 본 실험에서는 WSET과 2장에서 설명한 SSR과의 성능을 비교하였다. 먼저 표 2와 같은 샘플 데이터를 사용하여 SSR과 WSET로 유사도를 측정한 값을 비교하였다. WSET에 대해서는, SMM을 구축하기 위해 2개의 추상 클래스를 지정하였다. 표 2에 나타나는 웹 페이지들은 기본적으로 'programming'과 'travel', 두 개의 클래스로 분류된다. 웹 페이지 A, C, D, G, H, J는 'programming'에 관련된 주제를 다루고 있으며, 웹 페이지 B, E, F, I는 'travel'과 관련된 주제를 다루고 있다.

표 3과 표 4는 각각 SSR과 WSET으로 계산된 웹 페이지간의 유사도를 나타낸다. 표에서 A, B, ···, J는 웹 페이지를 나타내고, 각 칸의 값은 페이지 간의 유사도를 나타낸다. 실험 전에 예측한 바와 같이, 유사한 주제를 다루는 페이지 간의 유사도는 SSR의 경우 [0.135, 0.219], WSET의 경우 [0.216, 0.231] 사이의 값으로 비슷한 범위의 값을 가지는 것으로 나타났다. 하지만 유사하지 않은 페이지들임에도 불구하고 SSR에서는 유사한

표 2 Example data set

Page	Tag	Frequency	Page	Tag	Frequency
A	Java	25	F	Java	19
	Programming	13		Island	15
	Software	8		Tour	17
B	Java	13	G	Java	27
	Travel	21		Eclipse	15
	Island	5		Software	10
C	Eclipse	18	H	Eclipse	11
	Java	15		Programming	12
	Programming	13			
D	Software	18	I	Java	19
	Java	15		Indonesia	15
	Programming	13			
E	Travel	19	J	Java	19
	Island	15		Software	12
	Indonesia	8		Programming	7

표 3 SSR의 결과 값

Page	b	c	d	e	f	g	h	i	j
A	0.0521	0.179	0.199	0.0147	0.0840	0.186	0.159	0.113	0.211
B		0.0631	0.0892	0.143	0.127	0.0549	0.0177	0.117	0.102
C			0.157	0.0150	0.0942	0.156	0.242	0.129	0.135
D				0.0174	0.0921	0.138	0.164	0.122	0.219
E					0.130	0.002	0.002	0.101	0
F						0.0723	0.0123	0.1573	0.141
G							0.117	0.1051	0.210
H								0.0099	0.0250
I									0.211

표 4 WSET의 결과 값

Page	B	C	D	e	f	g	h	i	J
A	0	0.216	0.231	0	$1.1E^{-26}$	0.216	0.216	$1.4E^{-61}$	0.231
B		0	0	0.176	0.150	0	0	0.176	0
C			0.216	0	$1.1E^{-26}$	0.216	0.210	$1.3E^{-61}$	0.216
D				0	$1.1E^{-26}$	0.216	0.216	$1.4E^{-61}$	0.231
E					0.154	0	0	0.176	0
F						$1.0E^{-26}$	$1.1E^{-26}$	0.148	$5.1E^{-33}$
G							0.216	$1.2E^{-61}$	0.216
H								$1.3E^{-61}$	0.216
I									$1.1E^{-81}$

페이지로 판단되는 경우가 있었는데, 표에서 회색으로 칠해진 칸들이 그 예이다. 하나의 예로, 페이지 B와 J 간의 유사도를 보면 SSR는 두 페이지 간의 유사도를 0.102로 측정한 반면, WSET은 0으로 측정하였다. B와 J는 각각 다른 주제에 대한 페이지이지만, 두 페이지 모두 'java'라는 태그를 가지고 있다. 따라서 SSR은 두 페이지에 공통된 'java'라는 태그로 인하여 두 페이지가 유사하다고 측정하였고, 결과적으로 이것은 2장에서 설명한 바와 같이 SSR의 문제점을 나타낸다. 하지만 WSET에서는 모든 태그들이 같은 클래스에서 나타날 확률을 사용하여 웹 페이지 간의 유사도를 측정하기 때문에, SSR 보다 정확한 결과 값을 얻을 수 있다. 표 3과 표 4에서 회색으로 표시된 부분들은 모두 SSR의 이러한 문제점을 나타내는 부분이다.

4.2 실제 데이터

더욱 현실적이고 실제적인 결과를 얻기 위하여, http://delicious.com[6]에서 임의로 선택된 10,000개의 웹 페이지에 대한 태그 데이터를 이용하였다. 선택된 각각의 웹 페이지는 200번 이상 태그된 웹 페이지들이며, 10,000개의 웹 페이지에는 약 6,000개의 태그들이 부착되어 있었다. WSET을 적용하기 위해 50개의 클래스를 미리 정의하였다. 웹 페이지에 부착된 태그들 중에는 오타나 잘못된 값을 가지는 경우가 있었기 때문에, 실제 계산에는 각 웹 페이지에 부착된 태그들 중 가장 빈번하게 사용된 상위 60%의 태그들만을 사용하였다.

10,000개의 페이지 중, 유사하다고 판단되는 페이지들, 그리고 전혀 다르다고 판단되는 페이지들을 추출하여 이들에 대한 SSR과 WSET 값을 비교하였다. 표 5와 표 6은 유사하다고 판단된 페이지들과 그에 대한 SSR과 WSET 결과 값이다. 위의 페이지들은 서로 'design', 'portfolio' 등의 유사한 태그를 가지고 있었으며, 두 방법 모두에서 유사한 페이지라고 판단되었다.

표 7과 표 10은 서로 전혀 다르다고 판단된 페이지들과 그에 대한 SSR과 WSET 결과 값을 비교한 것이다. 모든 페이지들이 서로 공유하는 태그가 없으므로 SSR과 WSET 모두 0에 가까운 결과 값을 보여주었다.

마지막으로 표 8과 표 10은 둘 이상의 의미를 가지는 단어가 태그로 사용되었을 경우, SSR과WSET의 결과 값을 비교한 것이다. 이러한 다의어가 태그로 사용되었을 경우, SSR과 WSET은 서로 다른 결과 값을 나타내었다. 표 8을 보면 페이지 12와 13은 'webdev'라는 태그를, 페이지 14와 15는 'howto'라는 태그를 공유하고 있는 것을 알 수 있다. 페이지 12에서는 웹 디자인 쪽의 의미로서 'webdev'가 사용되었고, 페이지 13에서는 웹 프로그래밍 쪽의 의미로서 'webdev'가 사용되었다. 마찬가지로 페이지 14와 15에서 'howto'의 의미 또한 다르게 사용되었다. 페이지 14에서는 포토샵의 사용법을 나타내기 위해 사용되었고, 페이지 15에서는 윈도우를 사

표 5 Sample similar pages

	Web Pages	Tag Information
1	http://www.graphdrome.com/	[design, illustration, portfolio, ···]
2	http://inspiredology.com/graphic-design/typography	[typography, design, inspiration, font, ···]
3	http://feltron.com/	[design, portfolio, inspiration, typography, ···]
4	http://www.maxomatic.net/	[illustration, design, portfolio, graphic, ···]
5	http://www.adrianjohnson.org.uk/	[illustration, design, portfolio, ···]

표 6 Results of similar pages in 표 5

Page	2	3	4	5	2	3	4	5
1	0.017	0.018	0.069	0.038	0.016	0.015	0.011	0.042
2		0.031	0.021	0.027		0.030	0.023	0.016
3			0.018	0.028			0.046	0.015
4				0.038				0.011
	SSR Results				WSET Results			

표 7 Completely different pages

	Web Pages	Tag Information
6	http://developer.apple.com/tools/developonrailsleopard.html	[rails, ruby, osx, mac, development, ···]
7	http://www.overcomingbias.com/2008/02/my-favorite-lia.html	[education, teaching, learning, economics, ···]
8	http://www.photoattorney.com/	[photography, legal, law, copyright, ···]
9	http://www.sungevity.com/#start	[solar, energy, home, green, ···]
10	http://www.yumsugar.com/1663993	[coffee, recipe, food, dessert, cooking, ···]
11	http://www.chami.com/html-kit/services/favicon/	[favicon, webdesign, icon, tools, ···]

표 8 Pages with ambiguous tags

	Web Pages	Tag Information
12	http://960.gs/	[css, grid, webdesign, webdev, ···]
13	http://code.google.com/p/trimpath/	[javascript, framework, library, ajax, webdev, ···]
14	http://www.photoshopcafe.com/tutorials.htm	[photoshop, tutorial, graphics, howto, ···]
15	http://www.pctools.com/guides/registry/	[window, registry, tweak, howto, ···]

표 9 Results for ambiguous tags

	S_o(12, 13)	S_o(14, 15)
SSR	0.007	0.014
WSET	0	0

용하는 법의 의미로 사용되었다. 이 경우는 앞서 보인 유사한 페이지 또는 전혀 유사하지 않은 페이지들에 대한 결과와는 달리 두 방법이 다른 결과를 나타내었다. SSR을 통하여 유사도를 측정한 경우, 페이지 12와 13에 대해서는 0.007, 페이지 14와 15에 대해서는 0.014로 비교 대상이 된 페이지들이 서로 어느 정도 유사하다라는 결과 값을 나타낸 반면, WSET의 경우에는 두 경우 모두 0으로 전혀 유사하지 않다라는 결과 값을 나타내었다. 이 결과 값을 통해 샘플 데이터의 경우와 마찬가지로 SSR의 문제점을 확인할 수 있다. 즉, SSR은 두 웹 페이지에 속한 태그들 간의 유사도를 1:1로 비교하기 때문에, 두 페이지에서 'webdev'나 'howto'가 서로 다른 의미로 사용되었을지라도 각각 어떤 의미로 사용되었는지를 정확하게 구분하지 못하고 두 페이지가 유사하다고 판단한다. 반면에 WSET은 비교 대상이 되는 웹 페이지에 대한 태그 전체가 같은 클래스에서 나타날 확률을 이용하므로, 예를 들어, 페이지 14와 15의 경우 'howto', 'graphics', 'windows'가 같은 클래스에서 나타나야 페이지 14와 15가 유사하다고 판단하는 것이다. 하지만 위의 세 태그는 같은 클래스에서 나타나지 않으므로, WSET은 두 페이지를 유사하지 않은 페이지라 판단하고 올바른 결과 값을 나타낸다.

5. 고려 사항

본 논문에서 제안하는 방법인 WSET은 두 웹 페이지 간의 유사도를 측정하기 위해, 먼저 delicious.com과 같은 소셜 북마킹 시스템으로부터 웹 페이지-태그 데이터를 가져와서, 이 데이터로부터 SMM을 구축한다. 현실적으로 모든 웹 페이지-태그 데이터를 사용하여 SMM을 구축하는 데는 어려움이 있으므로 웹 페이지-태그 데이터에 대한 적절한 샘플링(sampling)이 필요하다. 이 때, 샘플링되는 웹 페이지-태그 데이터에 따라 편향된 결과를 가져다줄 가능성이 있으므로, 가능한 최신 데이터를 사용하여 SMM을 구축함으로써 최신 경향에 따르는 결과를 얻도록 해야 한다. 또한, 소셜 북마킹 시스템의 웹 페이지-태그 데이터는 실시간으로 변하는 한편, 본 논문에서 제안하는 방법은 특정 시점에 추출한 데이터로 구축된 SMM을 사용한다. 따라서 가능한 소셜 북마크 시스템에 새로이 추가된 최신 데이터를 반영하여 SMM을 주기적으로 구축해주어야 최신 데이터에 대한 비교 정보를 제공해 줄 수 있다.

마지막으로, 제안하는 방법은 SMM을 구축하기 위해 추상 클래스의 수 K를 사전에 지정해 주어야 한다. 이 때, 너무 적은 수의 K를 설정하면 유사하지 않은 태그들이 같은 추상 클래스로 분류될 가능성이 높아지고, 너무 많은 수의 K를 설정하면 유사한 태그들이 서로 다른 추상 클래스로 분류될 가능성이 많아진다. 따라서, 실제 비교 대상이 되는 웹 페이지들이 속할 주제들을 고려하여 적절한 수의 K를 지정해 주는 것이 필요하다.

6. 결 론

본 논문은 웹 페이지에 부착된 태그들을 사용하여 웹 페이지 간의 유사도를 측정하는 방법인 WSET을 제안하였다. 제안하는 방법은 먼저 Separable Mixture Model을 사용하여 주어진 웹 페이지와 태그들을 유사한 주제를 가지는 클래스들로 분류하고, 각 태그들이 각 클래스에서 나타날 확률을 구한다. 제안하는 방법은 이

표 10 Results of different pages in

Page	7	8	9	10	11	7	8	9	10	11
6	$7.8E^{-05}$	$6.1E^{-05}$	$1.5E^{-05}$	$1.6E^{-05}$	$8.9E^{-05}$	0	0	0	0	0
7		$1.5E^{-04}$	$6.7E^{-05}$	$4.2E^{-05}$	$2.0E^{-04}$		0	0	0	0
8			$7.6E^{-05}$	$6.2E^{-05}$	$6.2E^{-05}$			0	0	0
9				$1.5E^{-05}$	$1.0E^{-06}$				0	0
10					$8.4E^{-06}$					0
	SSR Results					WSET Results				

렇게 계산된 확률 값으로부터 두 웹 페이지가 가지고 있는 태그들이 같은 클래스에서 나타날 확률을 계산하고, 이를 기반으로 웹 페이지 간의 유사도를 측정한다. WSET은 여러 태그들을 가진 웹 페이지들을 비교할 때, 각 태그들을 일대일로 비교하는 것이 아니라 전체가 같은 클래스에서 나타날 확률을 기반으로 비교함으로써 웹 페이지에 부착된 태그들의 의미를 더욱 정확히 파악할 수 있다. 샘플 데이터와 실제 데이터를 사용한 실험 결과에 따르면 WSET은 두 가지 이상의 의미를 가지는 단어가 태그로 사용되는 경우, 기존 방법 보다 해당 태그의 의미를 잘 구분함으로써 보다 정확한 결과 값을 나타내었다.

참 고 문 헌

[1] Page L., Brin S., Motwani R., Winograd T., The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University Database Group (1998).

[2] J. M. Kleinberg: Authoritative Sources in a Hyperlinked Environment. In: 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, (1998)

[3] Shen X., Tan B., Zhai C., Implicit User Modeling for Personalized Search. In CIKM'05, ACM (2005).

[4] Chirita P., Nejdl W., Paui R., Kohlschutter C., Using ODP Metadata to Personalized Search. In Proc. of SIGIR (2005).

[5] Bao S., Xue G., Wu X., Yu Y., Fei B., Su Z., Optimizing Web Search Using Social Annotations. In WWW '07: Proceedings of the 16th International Conference on World Wide Web, pp.501-510, ACM (2007).

[6] Delicious social bookmarking, http://delicious.com/

[7] Hofmann T., Puzicha J., Statistical Models for Co-occurrence Data. Technical report, A.I.Memo 1635, MIT (1998).

[8] Wu X., Zhang L., Yu Y., Exploring Social Annotations for the Semantic Web. In WWW '06: Proceedings of the 15th International Conference on World Wide Web, pp.417-426, ACM (2006).

[9] Yanabe Y., Jatowt A., Nakamura S., Tanaka K., Can Social Bookmarking Enhance Search in the Web? In JCDL '07: Proceedings of the 2007 Conference on Digital Libraries, pp.107-116, ACM (2007).

[10] Heymann P., Koutrika G. Garcia-Molina H., Can Social Bookmarking Improve Web Search? In WSDM '08, ACM (2008).

[11] Law K., Harik G., Techniques for finding related hyperlinked documents using link-based analysis. U.S. Patent 6,754,873. June 22, 2004.

[12] Dean J., Henzinger M., Finding related pages in the World Wide Web. In Proc. of the Eighth International World Wide Web Conference (1999).

강 상 욱

2007년 2월 KAIST 전산학과 학사. 2009년 2월 KAIST 전산학과 석사. 2009년 2월~현재 삼성 DMC 연구소 소속. 관심분야는 Social Network/Web mining

이 기 용

정보과학회논문지 : 데이타베이스 제 37 권 제 1 호 참조

김 현 규

1997년 울산대학교 전산학과 학사. 2000년 울산대학교 전산학과 석사. 2000년~2001년 한국국방연구원 연구원. 2001년~2004년 LG전자 단말연구소 선임연구원 2005년~현재 한국과학기술원 전산학과 박사과정. 관심분야는 데이터베이스 시스템, 스트림 데이터 처리 등

김 명 호

정보과학회 논문지: 데이타베이스 제 37 권 제 1 호 참조