연관 웹 페이지 검색을 위한 코헤시브 아크 메저[1)]

이우기, 이병수○, 니디, 이원희
한인하대학교 산업공학과
trinity@inha.ac.kr, leebyoungsu@inhaian.net, dhini@inhaian.net, wonhee@inhaian.net

Cohesive Arc Measure for Web Navigation

Wookey Lee, Brian Lee○, Nidhi Rustagi Arora and WonHee Lee
Dept. of Industrial Engineering, Inha University

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 원하는 내용인지 알려면 웹 사이트를 하나하나 다시 탐색해야 할 것이다. 하지만 이제 검색엔진의 결과로 다량의 웹 페이지 리스트를 넘어서는 결과로써 좀 더 조직화되고 계층적인 결과가 요구되고 있다.

본 논문에서는 검색결과가 순서화된 하나의 리스트에서 벗어나 웹 사용자가 필요로 하는 구조화된 웹 페이지들의 집합을 보여주는 프레임웍을 제안한다. 이는 사용자 질의의 연관성을 판단하기 위해 웹의 내용과 구조 이 두 가지 모두를 사용하며, 여기서 말하는 구조란 같은 도메인이나 웹 사이트 안의 페이지들의 하이퍼링크로 연결된 구조를 의미한다. 웹의 각 페이지들의 구조는 논리적이며 개념적으로 서로 서로 관계되어 있다고 가정한다. 그러므로 웹 검색자가 입력한 키워드 질의를 반영해주는 논리적 구조를 가지고 있는 웹 페이지의 쌍을 찾아 작은 웹 그래프나 트리 구조로 보여주는 것이 본 연구의 목표이다.

본 논문에서 사용되는 기본 정의와 용어를 설명하고 시스템의 기술에 관하여 설명하겠다. 찾을 수 있는 모든 가능 키워드의 집합을 D 라고 표시하고, 검색자의 질의 Q는 키워드들의 집합 $<k_1,k_2,\ldots,k_n>$을 나타내며, W 는 키워드 k_i에 관련된 정보를 가지고 있는 웹 사이트 도메인의 집합을 의미 한다. 각각의 웹 사이트의 URL $W_i \in W$ 에서 W_i 의 도메인에 속해있는 노드들의 집합 즉, 웹 페이지들의 집합을 V_i 라하고 도메인 안의 웹 페이지들을 이어 주는 링크 즉, 하이퍼링크들의 집합을 E_i 라 할 때 유방향 그래프 G_i (V_i, E_i)라고 표현한다. V_i 집합은 같은 도메인 안의 웹 페이지만으로 제한한다. 다음으로 각각의 W_i가 가지는 최소 서브그래프를 $V_i^m \in V_i$ 이고 $E_i^m \in E_i$ 일 때 $G_i^m:(V_i^m,E_i^m)$라고 정의 한다.

본 논문에서는 최소 서브그래프의 노드집합 V_i^m안에 모든 웹페이지들의 유사도를 측정하기 위해 새로운 기법을 개발하였고 이것을 코헤시브 아크 메저(Cohesive Arc Measure)라고 한다.

코헤시브 아크 메저는 가중치를 구할 때 TF*IDF를 사용하며 이것은 텍스트 기반 데이터 분석의 수단으로 사용되는 방법으로 키워드의 숫자의 빈도를 이용하여 키워드와 페이지의 유사도를 측정한다. 코헤시브 아크 메저에서는 이 유사도를 이용하여 이번에는 페이지간의 유사도를 다시 산출하는 방법으로 이용한다. 웹 페이지 $p_j \in V_i^m$ 안에 있는 k 키워드의 $TF*IDF$ 를 다음과 같이 구한다.

$$c_k^j = tf(k)*idf(k)$$

페이지 pj $\in V_i^m$의 $f(p_j)$는 다음 식에서와 같이 페이지 안의 각 키워드의 $TF*IDF$값의 모음과 같다

$$f(p_j)=(c_2^j, c_3^j, \cdots, c_n^j)$$

1) 본 연구는 **지식경제부 및** 정보통신연구진흥원의 대학 IT연구센터 육성・지원사업 (**IITA-2008-C1090-0801-0031**)의 연구결과로 수행되었음.

이러한 *TF***IDF*값을 가지고 이번에는 코헤시브 아크 메저를 사용하여 링크로 연결되어있는 페이지들 간의 연관성을 다음과 같이 구한다.

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

두 가중치의 곱 ⊗는 다음과 같이 정의 된다.

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

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

$$= \frac{f(a)\cdot f(b)'}{2|f(a)||f(b)'|} + \frac{f(a)'\cdot f(b)}{2|f(a)'||f(b)|} = \frac{\sum_{x=2}^{n} c_x^a \times (1-c_x^b)}{2\sqrt{\sum_{x=2}^{n}(c_x^a)^2}\times\sqrt{\sum_{x=2}^{n}(1-c_x^b)^2}} + \frac{\sum_{x=2}^{n}(1-c_x^a)\times c_x^b}{2\sqrt{\sum_{x=2}^{n}(1-c_x^a)^2}\times\sqrt{\sum_{x=2}^{n}(c_x^b)^2}}$$

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 코헤시브 아크 메저를 사용하여 각 하이퍼 링크들의 가중치를 구하였다. 이 메저는 각 웹 페이지가 하이퍼 링크로 연결된 쌍들 중 키워드의 빈도수와 입력된 순서에 따라 검색자가 원하는 정보의 웹 서브 그래프를 찾아주었다.

참고문헌

[1] T. Phelps and R. Wilensky., "Robust Hyperlinks: Cheap, Everywhere, Now", *Digital Documents and Electronic Publishing*, LNCS2023, 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] D. Crabtree, P. Andreae, X. Gao, "Exploiting underrepresented query aspects for automatic query expansion", *SIGKDD*, pp.191-200,2007.

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

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

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

[8] K. Tajima, K. Hatano, T. Matsukura, R. Sano and K. Tanaka., "Discovery and Retrieval of Logical Information Units In Web", WOWS, pp. 13-23, 1999.

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

[10] http://directory.google.com/

[11] http://www.dmoz.org/

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

[13] G. Salton., *Introduction to modern information retrieval*, McGrawHill, 1987.

[14] D. Gusfield, *Algorithms on strings, trees and sequences*, Cambridge University press, 1998.

[15] M. A. Jaro., "Advances in record linkage methodology as applied to matching the 1985 census of Tampa", *Journal of American statistical association*, 84, 1989.