인용논문 분석을 통한 학술 문서 추천 시스템

박상진*, 김윤현**, 이지현**
*한국과학기술원 지식서비스공학
**한국과학기술원 문화기술대학원

e-mail: {psj9920, kimyoonhyun, jihyun lee}@kaist.ac.kr

A recommender of academic papers using the citation analysis

Sang Jin Park*, Yoon Hyun Kim**, Ji-Hyun Lee**
*Dept. of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology
**Dept. of Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology

요 약

인터넷의 급속한 보급으로 사용자가 정보와 지식의 접근이 용이 해진 반면, 방대한 정보의 과부화로 인하여 데이터의 신뢰성이 문제시 되고 있다. 특히, 기존의 학술 연구와 관련된 논문 데이터 검색에 있어서, 사용자의 요구 사항에 정확히 부합하는 결과물을 제공하는 데는 많은 한계를 가진다. 본 연구는 기존의 단순 키워드 매칭 검색의 한계를 넘어서, 레퍼런스와 인용 논문을 활용한 내용 기반 검색 방법론을 제안 한다.

1. 서론

최근 급속한 인터넷의 보급으로 인하여, 사용자의 정보와 지식 접근이 매우 용이 해진 반면, 매년 쏟아지는 정보 과부하로 인하여 그 신뢰성이 문제시 되고 있다[1]. 실제로 학술 분야에 있어서 2006 년에만 총 135 만 개가 넘는 저널 학술지가 편찬 되었다[2]. 따라서, 정보의 양은 방대해진 반면, 효율적인 인덱싱 작업이 데이터 증가량에 미치지 못하고 있는 상황이다. 이로 인해 사용자들은 정확히 자기의 요구 사항에 부합하는 데이터를 찾기가 매우 어려워 졌다. 이러한 문제를 해결 하고자 사용자의 검색을 도울 수 있는 추천 시스템의 연구가 활발히 진행되어 왔는데, 주로 데이터 마이닝 알고리즘을 활용한 방안들이 효율적으로 제시되고 있다[3,4]. 이 알고리즘은 수 많은 데이터들 사이에서 유용한 패턴들을 찾아내고 이를 분석하여 유저의 검색 쿼리와 가장 유사한 결과물을 도출하는데 이용 된다. 데이터 마이닝은 텍스트, 이미지, 오디오, 영상 데이터 등 여러 도메인에서 널리 쓰인다. 특히, 학술 저널 추천과 관련되어 Collaborative filtering[5], Content-based filtering [6] 이 두 가지 방안이 대중적으로 적용되어 왔으며 그 효율성 또한 널리 입증 되었다. 하지만, 이들 또한 데이터의 각 특징들에 대한 불충분한 명시화로 효율적인 검색에 있어서 여전히 한계를 지닌다 [7,8].
따라서, 본 연구는 기존의 한계를 citation analysis 를 통해 기존의 불충분 했던 각 학술 문서의 명시화를 보완, 확장 해주어 검색의 정확도를 높이고자 한다. 본 논문은 다음과 같이 구성된다. 2 장에서는 관련 연구에 대해 살펴보고, 3 장에서는 제안 방안에 대해 기술 하며, 마지막으로 4 장에서는 결론을 맺는다.

2. 관련 연구

기존의 방법들은 주로 문서들과 사용자 쿼리간의 유사성을 측정 후 높은 순서대로 추천을 해준다.
크게 대중적으로 두 가지 알고리즘이 이에 효율적으로 적용 된다.

2.1.1 Collaborative Filtering

Collaborative Filtering 은 주로 뉴스, 영화, 음악, 전자상거래 도메인에서 널리 이용 되고 있다. 자신과 비슷한 구매 패턴과 상품 평가 히스토리를 가진 사용자들을 파악하여, 이들이 높게 평가한 상품들을 나에게 재 추천 해준다. Collaborative Filtering 은 상품의 컨텐츠 분석을 필요로 하지 않기 때문에, 어떠한 도메인에서도 쉽게 적용 될 수 있다는 장점이 있다.
특히 전자 상거래 Amazon[9], e-bay[10] 등이 이를 적극적으로 반영하고 있다. 하지만, 초기에 사용자들의 적극적인 상품 평가를 필요로 하는 점과, 최근의 상품들에 대해서는 기존 상품 들 보다 추천 정확도가 떨어 지는 한계를 가진다[8].

2.1.2 Content-Based Filtering

Content Based Filtering 은 뉴스, Spam 메일 감지등과 같은 텍스트 분석에 탁월한 성능을 보여준다.
CBF 는 tf –idf [12]알고리즘을 활용하여 텍스트의 특징들을 Bag of words mode[11] 로 표현한다.
CBF 는 Collaborative filtering 과 달리 초기 조건들을 필요로 하지 않는 장점을 가진 반면에 도메인이 텍스트 기반이어야 한다는 것과, 정확도가 떨어지는 성능의 한계를 가진다[7].

3. 제안 방안

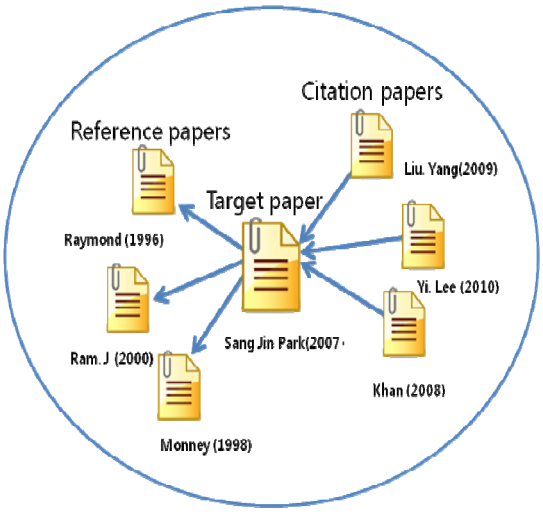

그림 1. 논문의 구조

그림 1 과 같이 하나의 대상 논문은 여러 레퍼런스와 인용 논문들을 가지고 있다. 우리는 이를 활용하여 기존의 Content-Based Filtering [6]을 보완, 개선 하고자 한다. 이는 각 문서들을 보다 심층적으로 분석 하여 사용자의 요구 사항에 적합한 추천을 해 줄 수 있게 해준다.

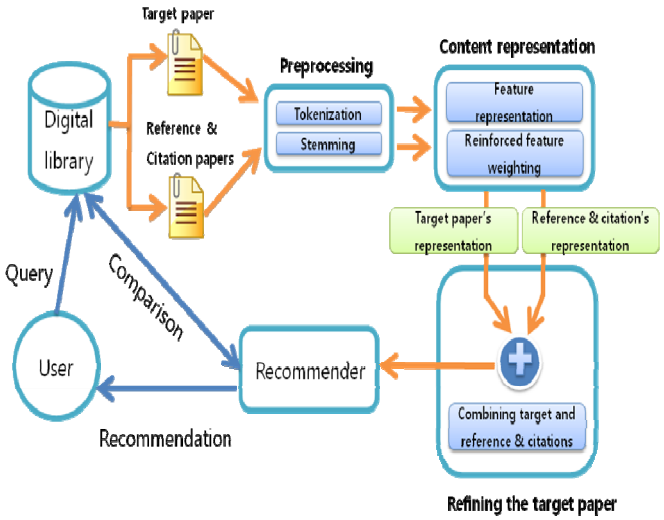

그림 2. 시스템 아키텍처

본 연구의 전체적인 아키텍처는 그림 2 와 같다.

3.1.1 Preprocessing:

1 차적으로 사용자의 쿼리와 가장 부합하는 대상 문서와 그와 관련된 레퍼런스, 인용논문들을 추출한다. 그 후 각 문서의 텍스트들은 stemming [13] 과정을 거쳐서 불용어를 제거 한다

3.1.2 Content representation:

각 문서의 특징들에 대한 명시화를 tf – idf [12] 를 이용하여 feature vector space 에 기술한다. 즉, 문서내의 각 단어 빈도수 tf 와 각 단어의 출현 해당 문서 빈도수 idf 를 곱한 값은 각 단어의 가중치가 되며, 이를 벡터 공간에 표현한다.
또한, 문서의 전반적인 내용을 함축 하고 있는 핵심 키워드들의 가중치를 올려 줌으로써, 문서 명시화 성능을 개선 해준다. 주로 논문의 제목이나 저자가 따로 분류한 키워드들이다[15]. 이는 다음과 같은 식 (1)로 가중치를 향상 시켜준다..

$$Enhanced\ t^{i}_{title} = idf\left(t^{i}_{title}\right) \times \left\{ tf\left(t^{i}_{title}\right) + tf - idf\left(t^{i}_{title}\right) \div 4 \right\}$$

$$Enhanced\ t^{i}_{keyword} = idf\left(t^{i}_{keyword}\right) \times \left\{ tf\left(t^{i}_{keyword}\right) + tf - idf\left(t^{i}_{keyword}\right) \div 4 \right\} \quad (1)$$

title 이나 keywords 에 해당되는 단어들에 한해, 기존의 tf –idf 값을 재조정 해준다.

3.1.3 Combing features of target and references & citations

대상 논문의 특징과 해당 레퍼런스, 인용 논문들의 특징들을 결합한다. 이에 앞서, 먼저 대상논문의 명시화를 보완 할 수 있는 유용한 논문들을 선출 하여야 한다. 모든 레퍼런스, 인용 논문들의 활용은 오히려 대상 논문 명시화에 방해가 된다. 이는 식(2)를 이용하여 선출 된다.

$$\text{If } (\text{Similarity}\left(f^{tgt}, f^{ref_i}\right) = \frac{f^{tgt} \cdot f^{ref_i}}{|f^{tgt}|\,|f^{ref_i}|} > \alpha)$$

$$\text{If } (\text{Similarity}\left(f^{tgt}, f^{cite_i}\right) = \frac{f^{tgt} \cdot f^{cite_i}}{|f^{tgt}|\,|f^{cite_i}|} > \alpha) \quad (2)$$

대상논문과 레퍼런스, 인용논문간의 코사인 유사성 측정을 통해서 일정한 기준 α 값 이상의 조건을 만족 시킬 시, 활용 가능 논문으로 분류 하였다. 본 연

구에서는 α 값을 0.87 로 정의 하였다.
또한, 식 (2)를 통해 선출된 논문들 일지라도 각 논문의 모든 특징벡터들을 다 활용 할 수는 없다. 각 논문에는 불필요한 특징벡터들 또한 다수 포함되어 있기 때문이다. 따라서, 유용한 특징벡터들 만을 선출하여야 한다. 이는 다음과 같은 식(3)을 통해 선별된다.

$$\text{useful } f^{ref_i} = \left(t_1^{ref_i}, t_2^{ref_i}, t_3^{ref_i}, t_4^{ref_i}, \cdots\cdots, t_m^{ref_i} \right)$$

$$\left(k = 1, 2, \cdots, m \quad , \quad t_k^{ref_i} > \alpha \right)$$

$$\text{useful } f^{cite_i} = \left(t_1^{cite_i}, t_2^{cite_i}, t_3^{cite_i}, t_4^{cite_i}, \cdots\cdots, t_m^{cite_i} \right)$$

$$\left(k = 1, 2, \cdots, m \quad , \quad t_k^{cite_i} > \alpha \right) \quad (3)$$

각 논문에서 단어의 가중치가 기준 값 α 이상의 조건을 만족 시킬 시, 활용 가능한 특징으로 분류 하였다. 본 연구에서는 α 값을 0.01 로 정의 하였다.
이 후, 각 레퍼런스, 인용 논문들이 대상 논문에 주는 영향력이 저마다 다르기 때문에 이점 또한 충분히 고려하여야 한다. 따라서, 우리는 각 레퍼런스, 인용 논문의 영향력 정도를 다음 식(4)와 같이 계산한다.

$$dis^{(tgt,\, ref_i)} = \text{similarity}\left(f^{tgt}, f^{ref_i} \right)$$

$$dis^{(tgt,\, cite_i)} = \text{similarity}\left(f^{tgt}, f^{cite_i} \right) \quad (4)$$

식 (2)를 통한 **similarity** 값을 이용하여 대상 논문과 각 레퍼런스, 인용논문들 간의 distance 를 구한다.
그 후. 레퍼런스, 인용논문을 활용하여 대상논문을 재명시화 해준다. 이는 식 (5)를 통해 구해진다.

$$\text{refined through reference } f^{tgt} = f^{tgt} + \sum_{i=1}^{m} dis^{(tgt,\, ref_i)} \text{ useful } f^{ref_i}$$

$$\text{refined through citation } f^{tgt} = f^{tgt} + \sum_{j=1}^{n} dis^{(tgt,\, cite_i)} \text{ useful } f^{cite_j} \quad (5)$$

$$\text{refined} f^{tgt} = \alpha \cdot \text{refined through reference } f^{tgt} + \beta \cdot \text{refined through citation } f^{tgt} \quad (6)$$

최종적으로 식(6)을 통해 레퍼런스를 통해 재구성된 대상논문과 인용논문을 통해 재구성된 대상 논문을 결합해 준다. 사용자가 레퍼런스와 인용 논문 사이에서 어디에 더 중요성을 두는지 여부에 따라서 α, β 값이 결정된다. 우리는 각각 0.7, 0.3 으로 정의 하였다. 이는 레퍼런스가 대상 논문에 더 많은 영향을 줄 것으로 판단되기 때문이다.

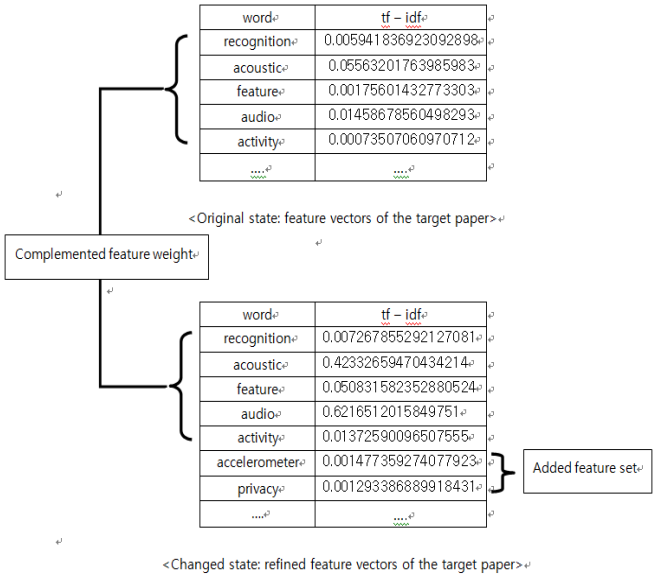

그림 3. 보완, 확장된 대상 논문의 특징

그림 3 에서 레퍼런스, 인용 논문을 활용한 후에 대상 논문의 각 특징들의 가중치가 보완되고, 또한 기존의 없었던 새로운 특징들이 추가 되면서 확장 되었음을 보여준다. .

3.1.3 Recommendation

레퍼런스, 인용 논문들을 활용하여 보완, 확장된 대상 논문과 디지털 라이브러리의 후보 논문들간의 유사성을 측정한다. 후에 이를 높은 순서로 5~7 개 후보 논문들을 사용자에게 추천을 한다.
유사성은 코사인 유사성 측정 방법을 통해 다음 식(7) 과 같이 계산된다.

$$\text{Similarity}\left(\text{refined} f^{tgt}, f^{cand_i} \right) = \frac{f^{tgt} \cdot f^{cand_i}}{\left| f^{tgt} \right| \left| f^{cand_i} \right|} \quad (7)$$

4. 결론

본 연구는 기존의 Content-based filering [6]방법의 단점이었던 불충분한 문서 명시화를 레퍼런스, 인용 문서들을 통해 보완, 확장 하였다. 이는 기존의 Link analysis 의 방법 중 PageRank [14]알고리즘에서도 적용

된 바 있지만, 단순 인용 횟수만을 이용한 추천이라는 점에서 한계가 있었다. 하지만 본 연구의 방법론은 문서 내용 자체의 심층적 분석을 통해서 추천을 해준다는 점에서 크게 근본적으로 다르다고 할 수 있다.

제안 알고리즘이 실제로 여타 다른 추천 기법들과 어느 정도의 성능차이를 개선 시킬 수 있는지 검증하는 것이 향후 과제이다.

참고문헌

[1] C. Shahabi, F.B.Kashani, J.Faruque(2001). A Reliable, Efficient, and Scalable System for Web Usage Data Acquisition. In: WebKDD'01Workshop in conjunction with the ACMSIGKDD 2001,San Francisco,CA,August

[2] Bjork, Bo-Christer; Roos, Annikki; Lauri, Mari Scientific Journal Publishing: Yearly Volume and Open Access Availability. Information Research: An International Electronic Journal, v14 n1 Paper 391 Mar 2009

[3] SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., AND TAN, P.-N. 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations 1, 2 (Jan.), 12–23.

[4] Raymond Kosala , Hendrik Blockeel, Web mining research: a survey, ACM SIGKDD Explorations Newsletter, v.2 n.1, p.1-15, June, 2000

[5] P. Resnick, N. Iacovou, M.Suchak, P. Bergstrom, J.Riedl.(1994) GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: Pages 175-186

[6] M.J. Pazzani and D.Billsus(2007). Content-Based Recommendation Systems.P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): The Adaptive Web, LNCS 4321, pp. 325 - 341, 2007

[7] M. Balabanović, Y. Shoham (1997). Fab: content-based, collaborative recommendation. Communications of the ACM CACM Homepage archive Volume 40 Issue 3, March 1997.

[8] M. Gr ˇ car, D. Mladeni ˇ c, B. Fortuna, and M. Grobelnik.(2005). Data Sparsity Issues in the Collaborative Filtering Framework. O. Nasraoui et al. (Eds.): WebKDD 2005, LNAI 4198, pp. 58- 76, 2006.

[9] http://www.amazon.com

[10] http://www.e-bay.com

[11] Z. Harris. (1985). Distributional Structure. Jerrold J. Katz (ed.)The Philosophy of Linguistics. Oxford University Press. 26–47

[12] G. Salton and C. Buckley. (1988). TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL. Information Processing and Management, v24 n5 p513-23 1988

[13] Porter, M.F(1980). An Algorithm for Suffix Stripping, Program, 14(3):130-137

[14] Y. Ding and E. Yan, A. Frazho, J. Caverlee (2009). PageRank for Ranking Authors in Co-citation Networks. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 60(11):2229–2243, 2009

[15] Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. Social Studies of Science 19, 473–496.CrossRef, Web of Science® Times Cited: 33