동사 레벨의 사용자 관심사 이해를 위한 오픈 디렉토리 프로젝트 확장 기법*

소슬기, 정다운, 류병걸, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail : {iminei, daounjung, smart123, yalphy}@korea.ac.kr

Open Directory Project Extension Scheme to Understand Verb-level User Interests

Seulgi So, Daoun Jung, Byung-Gul Ryu, SangKeun Lee
Division of Computer and Communication engineering, Korea University

요 약

개인화 서비스를 효과적으로 제공하기 위하여 사용자의 관심사를 이해하는 것은 매우 중요하다. 최근 많은 연구들이 사용자의 관심사를 이해하기 위하여 오픈 디렉토리 프로젝트를 이용하여 그 관심사를 주제별로 분류하는 방법을 제안하였다. 본 논문에서는 사용자의 관심사를 더욱 정확하게 이해하기 위하여 명사로 표현되어 있는 오픈 디렉토리 프로젝트를 동사 레벨로 확장하는 기법을 제안한다. 또한 실험 결과를 통하여 제안된 기법이 디렉토리와 연관된 동사를 효과적으로 확장하였음을 입증한다. 확장된 오픈 디렉토리 프로젝트는 사용자의 관심사를 동사 레벨로 이해하도록 함으로써 다양한 개인화 서비스에 활용될 것이다.

1. 서론

개인화 서비스는 각기 다른 관심사를 가진 사용자들에게 각 사용자에 적합한 서비스를 제공한다. 예를 들어 개인화 검색 엔진은 관심사가 다른 두 사용자가 동일한 쿼리를 요청한 경우, 각 사용자에게 다른 결과를 반환하여 준다. '카페'라는 쿼리의 경우 '온라인 커뮤니티'를 의미하는 동시에 '커피와 음료를 판매하는 장소'의 의미를 모두 가지고 있기 때문에 개인화 검색 엔진은 쿼리를 요청한 사용자의 관심사에 따라 다른 결과를 제공하거나 결과를 재정렬하여 반환한다. 이러한 개인화 서비스를 효과적으로 제공하기 위하여 사용자의 관심사를 이해하는 것은 매우 중요하다.

이를 위해, 기존 연구[4][5]에서는 사용자의 관심사를 오픈 디렉토리 프로젝트(ODP)[1]의 디렉토리로 표현하였다. ODP 는 웹 사이트들을 주제별로 분류하기 위하여 만들어진 계층구조의 웹 디렉토리로써, 100 만 개 이상의 디렉토리와 490 만개 이상의 웹 사이트를 트리 형태로 관리하고 있다. ODP 는 규모와 품질 면에서 사용자의 관심사를 나타내기에 유용한 소스이다.

본 논문에서는 명사의 형태로 표현되는 ODP 디렉토리를 동사 레벨로 확장하는 기법을 제안한다. 동사는 명사에 대한 사용자의 의도를 나타내어 사용자의 관심사를 더욱 명확하게 표현할 수 있기 때문이다. 예를 들어 두 사용자의 관심사가 동일하게 '자동차'라고 하더라도 구체적인 의도는 다를 수 있다. 한 사용자는 차를 구매하는 것을 원하고 다른 사용자는 차를 빌리는 것을 원할 수 있다. 이러한 상황에서 사용자의 관심사를 '사다' 또는 '빌리다'와 같이 동사 레벨로 이해할 수 있다면 개인화 서비스는 더욱 정확한 결과를 반환할 수 있을 것이다. 동사 레벨로 확장된 ODP 는 사용자의 관심사를 더 정확하게 이해하도록 함으로써 개인화 서비스의 성능을 향상시킬 것이다.

2. ODP 확장 기법

본 논문은 ODP 를 확장하기 위하여 각 디렉토리를 잘 표현할 수 있고 디렉토리와 연관된 동사들을 많이 포함할 것으로 기대되는 문서를 생성하고, 그 문서로부터 디렉토리와 연관된 동사를 추출한다. 본 논문은 그러한 문서를 생성하기 위하여 ODP body 를 활용한다. ODP body 란 디렉토리에 분류되어 있는 웹 페이지들을 의미한다. 이러한 웹 페이지들은 해당 디렉토리와 동일한 주제를 다루고 있기 때문에 그 주제와 연관된 동사들을 많이 포함하고 있다. 해당 문서는 특정 ODP 디렉토리에 분류된 웹 페이지들의 제목과 간략한 설명들을 결합하여 생성된다. 또한 문서가 디렉토리를 충분히 설명할 수 있도록 각 디렉토리의 하위 디렉토리에 분류된 웹 페이지들도 문서를 생성하는

* 본 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0077925)

데에 사용한다.

생성된 문서로부터 디렉토리와 연관성이 높은 동사를 추출하기 위하여, 본 논문은 문서에 포함된 모든 동사를 추출하고 각 동사마다 연관성의 정도를 의미하는 점수를 계산한다. 문서 내 모든 동사를 추출하기 위하여 먼저 문서를 문장 단위로 분할하고 디렉토리 이름을 포함하는 문장만 선별한다. 그리고 각 문장을 단어 단위로 분할하고 Stanford Natural Language Processing Tool[2]을 이용하여 분할된 단어의 품사를 태깅하여 동사로 태깅된 단어만 남긴다. 다음으로 Potter stemmer[3]를 사용하여 동사의 어근만 남겨 동일한 어근을 가진 동사를 동일한 동사로 인지할 수 있도록 한다. 마지막으로 대부분의 문서에 포함될 수 있는 일반적인 동사들을 불용어로 정의하여 그 동사들은 제외시킨다.

본 논문은 이러한 과정을 통하여 추출된 동사들에 대하여 문서에서 발생한 빈도와 문장 내 디렉토리명과의 거리 정보를 기반으로 연관성 점수를 계산한다. 즉 문서에서 발생 빈도가 높을수록, 문장 내 디렉토리명과의 거리가 가까울수록 디렉토리와 연관성이 높은 것으로 간주한다. 동사와 디렉토리 간의 연관성을 나타내는 *RelScore* 는 아래와 같이 계산한다.

$$RelScore(v_i, c) = \frac{nvf(v_i)}{avd(v_i, c)} \quad (1)$$

식 (1)에서 v_i 는 i 번째 동사를 나타내며, c 는 디렉토리명을 의미한다. $nvf(v_i)$는 정규화된 동사의 발생 빈도를 의미하여, $avd(v_i,c)$는 동사와 디렉토리명 사이의 평균 거리를 의미한다. $nvf(v_i)$는 아래와 같이 계산한다.

$$nvf(v_i) = \frac{vf(v_i)}{\sum_{w \in V} vf(w)} \quad (2)$$

식 (2)에서 $vf(v_i)$는 동사의 발생 빈도를 의미하며, V 는 문서 내에서 추출된 모든 동사의 집합을 의미한다. 즉 $nvf(v_i)$는 동사의 발생 빈도를 문서 내에서 발생한 모든 동사의 발생 빈도의 합으로 정규화한 값을 의미한다. 또한 $avd(v_i,c)$는 아래와 같이 계산한다.

$$avd(v_i, c) = \frac{1}{vf(v_i)} \sum_{l \in loc(v_i)} dist(l, c) \quad (3)$$

식 (3)에서 $loc(v_i)$는 동사 v_i 의 위치 집합을 나타낸다. 이는 문서 내에서 동일한 동사가 여러 번 등장하기 때문에 각 위치값을 얻기 위한 함수이다. 또한 $dist(l,c)$는 동사와 디렉토리명 사이의 거리값을 얻기 위한 함수이다. 즉 *RelScore* 는 정규화된 동사의 발생 빈도 값이 높을수록 동사와 디렉토리명 사이의 평균 거리 값이 낮을수록 높은 값을 갖는다. 모든 동사에 대하여 *RelScore* 를 계산한 후 가장 높은 연관성을 갖는 상위 K 개를 선택하여 ODP 디렉토리를 확장한다.

3. 성능 평가

본 논문에서는 제안 방법의 성능과 포함률을 검증하였다. 이를 위하여 먼저 기존 연구[6]에서 휴리스틱 룰을 사용하여 추출한 3,143 개의 ODP 디렉토리를 대상으로 제안 방법을 적용하였다. 제안 방법의 성능을 검증하기 위하여 정확도, 재현율, F-score 를 측정하였고, 포함률을 검증하기 위하여 3,143 개 디렉토리 중 5 개 이상의 동사를 추출한 디렉토리의 비율을 측정하였다. 성능을 측정하기 위하여 세 명의 자원자가 동사와 디렉토리 간 연관성을 판단하였고 두 명 이상의 자원자가 동사와 디렉토리 간 연관성이 있다고 판단된 동사를 정답으로 간주하였다.

<표 1> 동사 추출 방법의 성능과 포함률

정확도	재현율	F-score	포함률
0.76	0.46	0.57	86%

표 1 을 통하여 제안 방법이 높은 성능을 보장하는 것을 입증하였고 포함률은 86%로 높게 나타났다. 불용어를 적용하기 전의 정확도는 0.59 였지만 불용어 적용 이후 0.17 의 정확도가 증가하여 불용어의 중요성을 보여주었다.

제안 방법의 실증적 예제를 살펴보면 ODP 디렉토리 ‘ Business/consumer Goods and Services/Floral/Flower and Foliages’ 에 대하여 제안 방법을 통해 추출한 상위 5 개의 동사는 dry, preserve, stem, manufacture, air-dry 였다. 결과에서 볼 수 있는 것과 같이 ‘ 꽃과 잎’ 의 주제와 연관성이 높은 동사인 말리다, 보호하다, 줄기를 제거하다, 공기로 말리다 등의 동사가 추출되었음을 확인할 수 있다. 이를 통하여 본 논문은 제안한 기법이 각 ODP 디렉토리에 대하여 연관된 동사를 효과적으로 확장할 수 있었음을 입증하였다.

4. 결론 및 향후 연구

본 연구에서는 사용자의 관심사를 표현하기 위하여 ODP 디렉토리를 동사 레벨로 확장하는 기법을 제안하였다. 주어진 디렉토리를 ODP body 를 활용하여 문서로 표현하였고, 그 문서로부터 빈도와 거리를 기반한 연관성 정도를 측정하여 상위에 랭크된 동사를 추출하였다. 추출된 동사는 높은 성능과 포함률을 보였으며, 실증적 예제를 통하여 제안 기법이 연관된 동사를 추출하여 디렉토리를 효과적으로 확장함을 입증하였다. 향후, ODP body 뿐 아니라 디렉토리를 잘 표현할 수 있는 다양한 외부 소스를 활용한 연관 동사 추출을 통하여 ODP 확장 기법의 성능과 포함률을 개선하고, 불용어를 확장해 나갈 수 있도록 연구를 진행할 것이다.

참고문헌

[1] The open directory project, http://www.dmoz.org/.
[2] http://nlp.stanford.edu/software/tagger.shtml
[3] http://lucene.apache.org/java/docs/index.html
[4] M. Speretta and S. Gauch. Personalized search based on user search histories. In Web Intelligence, pages 622{628, September 2005.
[5] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In Proc. SIGIR '08, pages 155{162, July 2008.
[6] J. Ha, J.-H. Lee, K.-S. Shim, and S. Lee. Eui: An embedded engine for understanding user intents from mobile devices. In Proc. CIKM '10, pages 1935{1936, October 2010.