연관률 기반 복합어를 이용한 개선된 정보검색 시스템

이병희[O] 최종필** 박승규* 김민구**
아주대학교 정보통신 전문대학원[O*], 아주대학교 정보 및 컴퓨터 공학부**
{acecult[O], cjp**, sparky* }@ajou.ac.kr, minkoo@ajou.ac.kr**

Improved Information Retrieval System Using Multi word Based On Association Ratio

Byonghui Lee[O] , Seungkyu Park
Graduate School of Information and Communication, Ajou University,
Jongpil Choi, Minkoo Kim
College of Information & Computer Engineering, Ajou University

요 약

복합어의 추출은 정보 검색 및 온톨로지 분야의 연구에 있어서 중요한 비중을 차지하고 있다. 이 분야의 연구는 언어학적인 필터링 및 통계적 기법에 기반한 연구와 최근의 문맥 정보 및 사전 정보를 이용하는 기법 등으로 구분될 수 있다. 복합어를 정보 검색 및 온톨로지 분야에 응용하기 위해서는 복합어의 정확한 추출뿐만 아니라, 그 복합어가 문서를 표현할 수 있는 정도를 측정하는 기법이 필요하다. 특히, 정보검색 분야에서는 추출된 복합어에 대해 어떻게 가중치를 부여할 것인 가가 중요한 문제이다. 본 논문에서는 연관률(Association Ratio)에 기반하여 복합어를 추출하고, 추출된 복합어에 대해 적절한 가중치를 부여함으로써 검색 시스템의 성능을 향상 시킬 수 있는 방법을 제안한다.

1. 서 론

정보검색 분야에서 성능을 향상시키기 위한 연구는 활발히 진행되어 왔다. 문서간의 링크 정보를 이용하거나, 사전적 정보를 결합하는 등의 방법 등이 있었으며, 복합어를 검색에 이용하여 성능을 개선하고자 한 연구도 활발히 진행되었다[4]. 또한 지식을 형식화하고 이용하기 용이한 형태로 가공하기 위한 노력의 하나인 온톨로지 분야에서도 복합어의 정확한 추출은 중요한 연구 분야 중 하나이다. 따라서 복합어를 효율적으로 정확하게 추출하는 것은 정보검색 및 온톨로지 분야에 큰 도움이 될 수 있다.

복합어의 추출에 있어 어려운 점은 과연 추출된 복합어가 정확한 것인가에 대한 평가가 쉽지 않다는 것이다. 많은 연구들이 추출된 복합어의 목록을 전문가에게 의뢰하여 타당성을 부여하는 식의 평가 방법을 이용하고 있다. 본 논문에서는 추출된 복합어를 정보검색에 이용하여 그 성능을 평가하는 간접 평가 방법을 이용하였다. 정보검색에 복합어를 이용하기 위해서는 복합어의 정확한 추출뿐만 아니라 그 복합어에 가중치를 어떻게 부여할 것인가가 중요한 문제이다.

본 논문의 구성은 2장에서 관련 연구에 대한 분석을 살펴보고 3장에서는 기본적인 복합어 추출 기법과 추출된 복합에 가중치를 부여하는 방법에 대해 살펴볼 것이다. 4장에서는 실험 결과에 대해 살펴보고, 마지막으로 5장에서 결론 및 향후 과제에 대해 서술한다.

* 본 논문은 과학기술부의 국가지정연구실 사업(과제명:차세대 인터넷을 위한 지능형 온톨로지 자동생성 시스템 개발, 과제 번호:M10302000087-03J0000-04400) 지원으로 수행되었음

2. 관련 연구

복합어 추출에 대한 연구는 크게 문맥 정보에 기반한 방법과 통계적 정보를 이용하여 연관률을 측정하는 방법으로 구분할 수 있다[2][3]. 최근에는 언어학적 필터링 기법에 통계적 정보를 결합한 방법도 많이 연구되고 있다.

2.1 연관률(Association Ratio)에 기반한 연구

연관률에 기반한 연구는 단어들의 연관성을 계산하여 결합력을 측정하는 방법으로 두 개 이상의 단어가 결합할 수 있는 확률을 구하는 방법이다. 모든 단어들에 대해 연관률을 계산하는 것은 무리이기 때문에 복합어를 구성할 가능성이 높은 패턴을 이용하여 후보 복합어들을 추출하고 그에 대해 통계적 데이터를 이용하여 연관률을 계산할 수 있다[2]. 흔히 이용하는 통계적 데이터는 후보 복합어를 구성하는 각 단어의 출현 빈도와, 후보 복합어를 구성하는 모든 단어가 함께 출현한 빈도, 그리고 모두 출현하지 않은 빈도 등이다. 이러한 데이터를 이용하여 몇 가지 계산 방법을 이용하여 그 연관률을 측정한다.

2.2 문맥 정보에 기반한 연구

문맥 정보에 기반한다고 하여 통계적 정보가 전혀 이용되지 않는 것은 아니며, 통계적 데이터로서 C-Value가 많이 이용된다[3]. 단어의 출현 빈도수에 의해 민감한 영향을 받는 C-Value는 해당 단어가 다른 복합어의 부분집합일 경우 그 가중치를 낮추게 된다. 이는 상대적으로 다

른 복합어를 포함하는 단어에 대해 그 중요성을 높여주기 위한 것이다. 이러한 기초 데이터를 기반으로 문맥 정보를 이용하는데, 문맥으로서의 가치를 평가하는 방법으로는 NC-Value가 있다. NC-Value는 단어가 얼마나 많은 복합어 또는 다른 단어들과 함께 나왔는지의 정도를 측정하는 방법이다. 이외에도 의미망을 이용하는 SNC-Value가 있다. SNC-Value에서는 단어가 해당 복합어와 사전적으로 얼마나 가까운지를 UMLS 등의 의미망을 이용하여 측정한다[3].

2.3 정보검색 분야에서의 복합어의 이용

정보검색 분야에서는 문맥 정보에 기반한 복합어의 추출 기법보다는 상호 정보(Mutual Information)에 가까운 연관률 기반의 복합어 추출 기법이 보다 많이 이용되고 있다[4]. 문맥 정보에 기반한 복합어의 경우 추출에 걸리는 시간이 연관률 기반 방법에 비해 오래 걸리고, 추출된 복합어가 일반적인 의미의 복합어인 경우가 많다. 따라서 정보검색에 복합어를 이용하는 경우 연관률에 기반하여 보다 중요한 가치를 가지는 복합어를 추출하여 문서의 특징으로 이용하거나 질의어를 확장하는 등의 방식을 취하고 있다.

3. 연관률 기반 복합어 추출 및 가중치 부여

3.1 연관률 기반 복합어 추출

본 논문에서는 복합어를 연관률에 기반하여 추출하였다. 연관률은 상호 정보에 가까운 방법으로 비교적 수식이 간단하고, 상호 함께 출현할 확률이 높은 단어들을 결합하는 방식으로 복합어를 추출한다. 복합어의 연관률을 계산하기 위해서는 우선 후보 복합어를 추출해야 한다. 일반적으로 후보가 될 복합어를 추출하기 위해서는 단어의 품사정보를 이용하여 패턴을 만들어 그 패턴에 일치하는 단어들을 후보 복합어로 추출한다. 본 논문에서는 형용사 + 명사, 명사 + 명사의 두 가지 패턴을 이용하여 후보 복합어를 추출하였다. [1]에 따르면 위의 패턴을 이용하는 것이 정확도는 높여주지만 재현율은 감소시킬 수 있다고 지적하고 있다. 이외에도 전치사를 추가하거나 좀더 세부적인 패턴을 지정하여 후보 복합어를 추출하는 방법도 있다.

본 논문에서는 [1]에서 이용한 수식을 이용하여 복합어의 연관률을 계산하였다. 연관률 계산을 위해 이용되는 통계적 데이터는 다음과 같다. 복합어 L이 i, j의 두 단어로 구성되었다고 할 경우,

a	i와 j가 함께 출현한 빈도수
b	i만 출현한 복합어의 빈도수
c	j만 출현한 복합어의 빈도수
d	i와 j가 모두 출현하지 않는 복합어의 빈도수

표1. 연관률 측정을 위한 통계적 데이터

위와 같은 통계적 데이터를 이용하여 다음과 같은 세 가지의 연관률 계산 방법을 적용하였다.

$$IM = \log 2 \frac{a}{(a+b)(a+c)}$$

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

$$Loglike = a\log a + b\log b + c\log c + d\log d - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + (a+b+c+d)\log(a+b+c+d)$$

또한, 추출된 후보 복합어에 대해 더 이상의 복합어가 추출되지 않을 때까지 반복하여 중첩된 복합어에 대한 추출도 가능하다. 예를 들어 (t1, t2, t3)가 모두 명사일 경우 (t1, t2), (t2, t3)의 조합이 가능한데, 각 후보 복합어의 연관률을 계산하여 보다 높은 확률을 가지는 후보를 복합어로 인정하여 (t1, t2)가 t4라는 명사로 지정하면, 다음 주기에서 t4, t3의 조합이 가능해진다.

3.2 복합어에 대한 가중치 부여

정보 검색에서 복합어의 추출과 함께 중요한 것은 추출된 복합어에 대해 어떻게 가중치를 부여할 것인가이다. 본 논문에서 적용한 정보검색 모델은 랭귀지 모델인데, 문서의 길이와 출현 빈도를 이용하여 가중치를 계산한다[5]. 복합어의 경우 단일어에 비해 출현빈도가 적은 경향이 있고, 출현빈도가 높다 하더라도 일반적인 의미의 복합어일 가능성이 있어 적절한 가중치의 부여가 쉽지 않다. 따라서 본 논문에서는 복합어의 tf*idf 가중치를 계산하여 원래 문서집합에 추가하는 방법을 이용하였다. 문서집합에 추가하는 과정에서는 순위화된 복합어의 연관률이 평균 이상의 확률을 가질 경우에만 추가하도록 하였다. 이러한 방법을 이용하면 복합어가 문서를 표현하는 중요도를 측정할 수 있고 역문헌 빈도수를 이용하여 일반적인 의미를 가지는 복합어의 가중치를 떨어뜨릴 수 있는 장점이 있으며, 평균 이상의 연관률을 가지는 복합어만 추가함으로써 필터링의 효과를 거둘 수 있다.

추출된 복합어를 필터링 없이 추가할 경우 오히려 검색 성능을 악화시킬 수 있고, 대부분의 복합어 추출에 관한 연구에서도 필터링을 통해 중요하지 않은 복합어를 거르는 과정을 거치고 있다. 본 논문에서는 시스템의 효율성을 위해 필터링 과정을 가중치 부여과정에 통합하여 수행하였다. 복합어에 대해 tf*idf 가중치를 부여하여 랭귀지 모델에 적용함으로써 복합어가 정보검색에 미칠 수 있는 영향을 실험을 통해 알아보았다.

4. 실험 및 결과

실험은 NIST에서 주관하는 TREC 컨퍼런스의 데이터 중 AP 기사 (88, 89, 90년) 모음을 이용하였다. AP 기사 모음은 기사의 특성 상 비교적 많은 복합어를 포함하고 있어 실험에 용이한 문서집합이라 할 수 있다. 실험을 위해 설계 구현한 시스템의 개략적인 구성도는 다음과 같다.

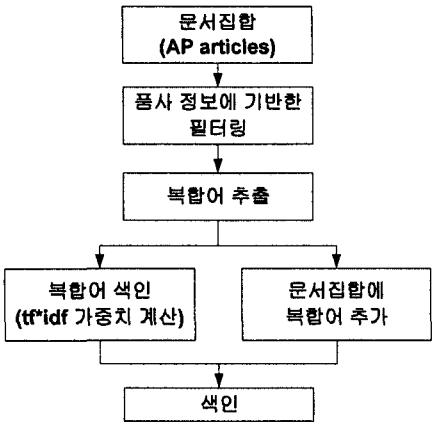

그림 1. 복합어를 이용한 검색 시스템

본 실험을 위한 테스트 질의어는 TREC8 데이터를 위한 질의어의 집합 중 복합어를 포함하고 있는 50개의 질의어를 선정하여 이용하였다. 본 실험은 복합어를 검색에 이용할 경우 얻을 수 있는 효과를 확인하고자 하는 것이므로 이미 추출된 바 있는 복합어를 포함하고 있는 질의어에 한정하여 실험을 수행하였다. 선정된 질의어는 평균 1.6개의 복합어를 포함하고 있었으며, 실험에서는 질의어에 추출된 복합어를 추가하여 이용하였다.

테스트 문서집합에 대한 정보를 살펴보면 다음과 같다.

문서의 개수		242,803
단일어의 개수		219,476
복합어의 개수	Loglike	556,925
	IM	513,688
	ϕ^2	481,941
	총 개수	1,426,686

표2. 테스트 데이터 관련 정보

위의 표에서 총 개수는 품사 정보에 기반하여 만들어진 패턴(형용사 + 명사, 명사 + 명사)에 의해 추출된 총 후보 복합어의 개수를 나타낸다. 각 수식에 의해 추출된 복합어에 대해 문서집합에 추가하여 색인 하는 과정에서 해당 복합어의 가중치는 복합어만으로 구성된 문서집합에서 구한 tf*idf 가중치를 부여하였다.

앞서 설명한 바와 같이 평균 이상의 연관률을 가지는 복합어만 문서집합에 추가하도록 하였고, 평균적으로 30% 정도의 복합어가 문서집합에 추가되었다. 검색은 랭귀지 모델에 기반하여 수행하였고, 검색 결과를 보면 다음과 같다. 여기서 10 docs, 20 docs는 상위 10개, 20개 문서에서의 정확률을 나타낸다. Average는 전체 검색 결과에서의 평균 정확률을 의미한다. 검색 결과 Loglike와 ϕ^2의 경우 Baseline에 비해 우수한 결과를 보였고, IM의 경우

	10 docs	20 docs	Average
Baseline	0.34	0.326	0.259456
Loglike	0.358	0.34	0.273003
ϕ^2	0.35	0.329	0.271032
IM	0.253	0.256	0.231958

표3. 검색 결과

다소 낮은 성능을 보였다. 이는 IM의 경우 두 단어가 함께 출현한 빈도수가 미치는 영향이 다른 수식의 경우에 비해 크기 때문으로 분석된다. 그 결과 일반적인 의미를 가지는 복합어가 상위에 위치할 가능성이 높아 복합어가 문서를 표현하기에 적합한지의 평가가 제대로 이루어지지 않았던 것으로 풀이된다. Loglike와 ϕ^2의 경우 비교적 일반적 의미를 가지는 복합어가 하위에 위치하여 Baseline보다 좋은 결과를 유도한 것으로 보인다.

5. 결론 및 향후 과제

본 연구에서는 앞서 언급한 연관률 기반의 복합어 추출 기법을 정보검색에 이용함으로써 그 성능을 향상 시킬 수 있는 방안을 소개하였다. 연관률 기반의 복합어 추출 기법은 다른 연구에 비해 과정이 비교적 간결하고 수행시간을 절약할 수 있으며, 복합어의 전문성을 측정하기에 용이한 장점이 있다. 추출된 복합어를 정보검색에 이용하기 위해, 본 연구에서는 복합어 만으로 구성된 문서집합을 별도로 생성하여 복합어의 tf*idf 가중치를 구해 검색에 이용하는 방법을 이용하였다. 실험 결과, 복합어를 이용하지 않는 경우에 비해 비교적 우수한 성능을 얻을 수 있었다. 또한, 복합어의 연관률을 계산하는 방법에 따라 검색에 미치는 영향이 달라지는 것도 알 수 있었다. 향후, 복합어의 필터링 과정에 대해 단순히 평균 이상이라는 조건이 아닌 보다 정교한 조건을 연구하고, 추출된 복합어에 대한 가중치 부여 방법을 개선하여 연구를 진행할 계획이다.

6. 참고 문헌

[1] Daille, B., Study and implementation of combined techniques for automatic extraction of terminology. In The Balancing Act: Combining Symbolic and Statistical Approaches to Language. The MIT Press, 1996

[2] Kenneth Ward Church, Patrick Hanks, Word Association Norms, Mutual Information, and Lexicography, Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics , 1989

[3] Katerina T. Frantzi, Sophia Ananiadou, Junichi Tsujii: The C-Value/NC-Value Method of Automatic Recognition for Multi-word Terms, ECDL, 1998

[4] John Bear and David Martin, Using Information Extraction to Improve Document Retrieval, Text Retrieval Conference. 1996

[5] Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to In Information Retrieval. In 21th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.