다중단어를 사용한 정보검색 시스템에서의 재현정확도 향상방법

최종희[1], 최동시[1], 박세영[1], 오희국[1]
한양대학교 전자계산학과[1], 한국전자통신연구원[1]

A Method for Improving Recall Precision on Information Retrieval Systems Using Multiple Terms

Jonghee Choi[1], Dongsi Choi[1], Seyoung Park[1], Heekuck Oh[1]
Dept of Computer Science, Hanyang Univ[1], ETRI[1]

요 약

정확한 정보를 검색하기 위해 단일단어를 사용하는 대신에 다중단어를 사용하는 정보검색 시스템에 대한 연구가 활발히 진행되고 있다. 그러나 아직까지 다중단어를 이용한 검색시스템은 그리 많지 않다. 다중단어를 이용한 정보검색시스템의 한 예가 키팩트를 이용한 정보검색 시스템이다. 키팩트란 키워드뿐만 아니라 관련정보를 같이 포함하고 있는 다중단어의 하나다. 키팩트에 기반한 정보검색 시스템은 현재 문서의 색인과정과 질의어의 키팩트 추출과정에서 같은 가중치를 가진 키팩트를 생성한다. 그러나, 하나의 명사구는 그것이 갖는 의미에 따라 각기 다른 다양한 키팩트를 생성하기 때문에, 이들의 결과에 기존의 정보검색 방법을 적용하는 것은 문제가 많다. 따라서 본 논문에서는 색인시에 생성되는 각각의 키팩트에 적절한 가중치를 부여함으로써 보다 정확한 정보검색이 이루어지도록 하는 방법을 제안한다.

1. 개 요

대부분의 정보검색 시스템은 단일어인 키워드를 사용하여 정보를 검색한다. 키워드로 정보를 검색할 경우에 찾고자 하는 정보가 모호성을 갖게 된다. 모호성이란 하나의 단어가 여러 가지 의미를 가지고 있거나, 너무 광범위한 의미를 가지고 있는 것을 말한다. 이런 모호성을 해소하기 위한 한가지 방법이 다중단어를 사용하는 것이다. 다중단어는 기존의 정보검색 시스템이 사용하고 있는 키워드라는 단일어와는 달리 이 키워드와 관련된 정보를 같이 포함하고 있다. 다중단어는 관련정보를 같이 포함하고 있기 때문에 그 키워드가 갖는 의미를 정확하게 파악할 수 있다. 관련정보란 키워드가 가지는 의미를 정확히 알 수 있도록, 키워드의 특성을 설명해 주는 정보를 말한다. 그래서, 복합명사로 구성된 경우는 각이 동반되는 명사가, 관형어를 갖는 경우는 관형어가, 그리고 문장의 형태로 나타나는 경우는 동사나 형용사가 관련정보의 역할을 한다 [1-3]. 이렇게 해서 다중단어의 개념으로 나온 것이 키팩트이다.

본 논문에서는 키팩트를 이용한 정보검색 시스템에서 보다 정확한 정보를 검색하기 위한 방법으로 각각의 키팩트에 가중치를 부여하는 방법을 제안한다. 그리고 이렇게 했을 때의 정확도와 동일한 가중치를 부여했을 때의 정확도를 비교하였다. 2장에서는 키팩트에 대한 설명을 하고, 3장에서는 키팩트를 추출하는 방법에 대해서 설명한다. 4장에서는 색인시에 어떻게 가중치를 부여하는지에 대해 설명하고, 5장에서는 각각의 키팩트에 다른 가중치를 부여했을 경우와, 동일한 가중치를 부여했을 경우에 검색된 결과의 성확도를 실험을 통해 비교한다. 그리고, 마지막으로 6장에서는 결론 및 향후 연구방향에 대해 설명한다.

2. 키팩트

정보를 검색할 때, 단순히 키워드뿐만 아니라 관련된 정보를 같이 가지고 있는 키워드로 검색할 경우 검색의 정확성을 높일 수 있다. 이것은 검색하고자 하는 범위가 축소되기 때문이다. 사용자가 하나의 단어로 질의문을 구성하기 않고 여러 개의 단어로 질의문을 만드는 것은 바로 정확한 의미를 나타내기 위해서이다.

키팩트란 문서의 내용을 대표하는 것이 단어(word)가 아니라 사실(fact)이어야 한다는 점에서 만들어진 말로서, 키팩트는 키워드와 관련정보를 가지고 있다. 키팩트는 중심어와 종속어로 구성되어 있는데, 키워드는 중심어, 관련정보는 종속어가 된다. 문장에서 표현방법은 여러 가지 이지만 그것이 나타내는 내용이 의미적으로 동일하다면 같은 키팩트라고 할 수 있다. 그래서 하나의 키팩트는 의미적으로는 같다 할지라도 문법적으로는 서로 다를 수 있다. 왜냐하면 하나의 키팩트를 표현하는 데는 여러 가지 형태와 표현방법이 존재할 수 있기 때문이다. 어떤 문서에서 기존의 방법으로 키워드를 추출하여 놓고 그 키워드만으로 원래의 문서를 추론하는 것과, 명사구를 추출한 후 그 명사구만으로 원래의 문서를 추론하는 것을 시험해 본 결과, 후자의 방법이 원래의 문서를 더 잘 표현하고 있다는 것이 증명되었다 [4].

색인어의 조건은 첫째, 문서를 대표하여야 하고 둘째 다시 나타날 확률이 있어야 한다는 것이다. 그러나, 보통의 명사구는 어느 정도 문

서를 대표하기는 하나 다시 나타날 확률은 거의 없다. 그래서 키팩트 기반 정보검색시스템에서 하나의 명사구는 여러 개의 키팩트로 생성되어야 한다.

3. 키팩트 추출방법

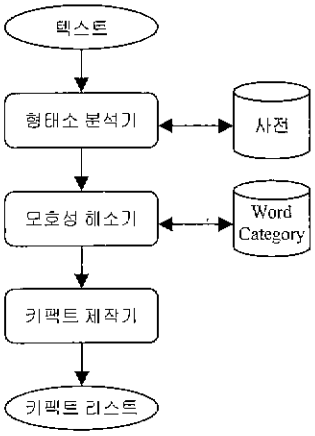

그림 1 키팩트 추출기 구성도

키팩트를 추출하기 위해서는 크게 세 단계의 과정을 거쳐야 한다. 일단 주어진 문장이나 어절을 형태소로 분석해야 하고, 두 번째로 분석된 형태소에서 모호성을 해소해야 한다. 모호성을 해소할 때는 같은 문장이나 같은 어절 내에 포함되어 있는 다른 형태소와의 상호 관련정도를 비교하여 모호성을 해소한다. 이때 쉽게 할 수 있는 방법이 관련명사를 사용하는 것이다. 그러나, 모호성을 완벽하게 해소하기에는 어려움이 많다. 그래서, 이 단계의 모호성 해소는 이주 간단한 패턴들에 대해서만 적용하고, 나머지들은 말뭉치에 나타나 있는 빈도수에 따라 결정한다. 그리고, 마지막으로 모호성이 해소된 형태소를 가지고 키팩트 생성규칙에 따라 키팩트를 추출한다.

형태소 분석기와 모호성 해소기를 거친 형태소들은 하나의 품사와 하나의 의미를 가진다. 키팩트를 추출하는 방법은 다음과 같다 [5]

- 중심어만으로도 키팩트가 될 수 있다. 즉, 하나의 명사(기존의 키워드)도 키팩트로 사용한다.
- 두 키워드가 '의'로 연결된 경우에 두 키워드는 하나의 중심어를 형성할 수 있으며, '의'의 다음 키워드는 종속어가 될 수 있다.
- 두 키워드가 '와/과'로 연결된 경우에 두 키워드는 하나의 중심어를 형성할 수 있으며, '와/과'의 다음 키워드는 종속어가 될 수 있다. 이때, 중심어와 종속어의 위치가 서로 바뀌어도 상관이 없다.
- 피생관형사, 서술격동사, 비서술격 동사는 관련동사로서 종속어로만 쓰일 수 있다.
- 조사 없이 연결된 두 키워드가 하나의 키워드를 형성하는 경우에는 순서를 갖는다.

키팩트를 만들 때, 하나의 명사구를 하나의 키팩트로 만드는 것이 아니다. 하나의 명사구에서 여러 개의 키팩트를 만들어 낸다. 이것은 명사구를 하나의 키팩트로 표현하는 것은 가능하지만, 다른 패턴에 의해 생성되는 키팩트와의 부분 매칭에 문제점이 있기 때문이다. 이렇게 생성된 키팩트 중에는 중심어와 종속어를 모두 가진 것도 있지만, 중심어만을 가진 것도 있다. 만일 이 키팩트를 가지고 검색을 하는 경우에 중심어와 종속어를 모두 가지고 있는 것과 중심어만을 가지고 있는 것에 같은 가중치를 부여한다면 보다 정확한 검색을 할 수 없을 것이다.

4. 색인과정

본 논문에서 제안하는 방법이 다른 기존의 정보검색 시스템과 크게 다른 점은 색인 과정에서 각각의 키팩트가 자신의 고유한 가중치를 갖고 있다는 점이다. 한 명사구에서 여러 개의 키팩트가 생성되기 때문에 모든 키팩트의 가중치를 같게 한다는 것은 문제가 있다. 예를 들어, "감상의 본질"이라는 명사구가 있다고 하자. 그러면 이 어절에서 다음과 같은 키팩트가 만들어질 것이다.

ex) [감상,NIL], [본질,NIL], [감상, 본질], [감상 본질, NIL]

이 경우에 [감상, NIL]과 [본질, NIL]보다는 [감상, 본질]이 더 정확한 의미를 갖는다. 그러므로, [감상, NIL], [본질,NIL]보다는 [감상, 본질]이 더 큰 가중치를 가져야 한다.

또 다른 예를 들어보면 "신과 자연의 질서"라는 명사구가 있다고 할 때 추출되는 키팩트는 다음과 같다.

ex) [신, NIL], [자연, NIL], [질서, NIL], [신, 질서], [자연, 질서], [신 자연,], [자연 신,], [신 자연, 질서],[자연 신, 질서]

이 경우에도 중심어만 있는 것보다는 중심어와 종속어를 함께 가지고 있는 키팩트의 가중치가 더 높아야 한다.

또, 위 두 가지 예에서 보듯이 각각의 키워드 "감상", "본질", "신", "자연" 그리고, "질서"는 각각 1번씩 문서의 본문에서 나타난 것이므로 이들로 인해 생성된 키팩트 내에서의 각 명사에 대한 가중치 합은 같아야 할 것이다. 다시 말해서, "감상의 본질"이라는 명사구에서 추출된 키팩트 중에서 "감상"은 전체 4개의 키팩트중 3개의 키팩트에서 출현했고, "신과 자연의 질서"라는 명사구에서 추출된 키팩트 중에서 "신"은 전체 9개의 키팩트중 6개의 키팩트에 출현했다. 그러므로, "감상"이 출현한 각각의 키팩트는 1/3에 해당하는 가중치를 부여받아야 하고, "신"이 출현한 각각의 키팩트는 1/6에 해당하는 가중치를 부여받아야 한다. 그러나, 더 정확한 값을 계산하기 위해서는 많은 실험과 고찰을 필요로 한다.

5. 검색 및 실험

모든 키팩트는 [중심어, 종속어]이거나 [중심어, NIL]이다. [중심어, 종속어]는 [중심어, NIL]보다는 더 협소하고 정확한 의미를 가지므로, 적합한 문서나 정보를 찾는데 더 큰 도움을 줄 것이다. 본 실험에서는 계몽사 백과사전 23,112개의 문서를 대상으로 하였고, 총 데

이터의 크기는 약 12Mbytes이다. 그리고, 실험은 두 가지의 경우로 나누어서 했다. 즉, 모든 키팩트가 동일하게 가중치를 갖는 경우와, 각각의 키팩트가 서로 다른 가중치를 갖는 경우로 나누어서 실험했다. 그리고, 후자의 경우에 다음의 공식을 이용했다

$$Weight = \frac{k}{N} \times p \qquad 식\ 1$$

식 1에서 N은 하나의 명사구에서 만들어진 전체 키팩트의 개수이고, k는 특정단어를 포함하고 있는 키팩트의 개수이다. 그리고, p는 상관계수로 본 논문에서는 중심어와 종속어가 모두 존재하는 경우에는 1을, 중심이만 존재하는 경우에는 0.5를 사용했다

이 두 가지 경우에 실제로 검색되는 문서는 일치한다. 같은 질의를 했을 경우에 같은 문서가 검색되지만, 두 경우의 순위는 서로 다르다. 순위부여 알고리즘은 벡터공간모델을 이용했다. 그리고, 정확도를 측정하기 위해 미리 질의에 대한 응답을 정의해 놓았다. 정확도는 검색된 결과가 미리 정의해 놓은 질의 응답과 얼마나 일치하는가를 나타낸 것이다. 그러나, 기존의 정확도의 개념을 약간 확장하여 사용하였다. 검색된 결과의 순위를 중요시하여 정확도를 결정하였다. 즉, 검색된 문서의 순위 상위 15위 내에 적합한 문서 몇 개가 포함되어 있는지를 비교하였다. 예를 들어, 미리 정의해 놓은 응답은 10개의 문서가 있다고 하자. 검색된 문서의 순위 상위 15위 내에 10개의 문서가 모두 들어있으면 정확도는 100%가 되고, 5개가 들어있으면 50%, 하나도 들어있지 않으면 0%가 된다.

다음은 몇 가지 질의에 대한 응답을 비교하고 있다.

1) "추석의 기원은?"

	키워드	동일한 가중치를 갖는 키팩트	다른 가중치를 갖는 키팩트
검색된 문서	466	314	314
정확도	75%	75%	75%

2) "장영실의 업적은?"

	키워드	동일한 가중치를 갖는 키팩트	다른 가중치를 갖는 키팩트
검색된 문서	352	258	258
정확도	77%	88%	88%

3) "지외선과 적외선의 차이는?"

	키워드	동일한 가중치를 갖는 키팩트	다른 가중치를 갖는 키팩트
검색된 문서	1075	738	738
정확도	30%	40%	60%

첫 번째 예에서는 세 가지 경우에 대해 같은 정확도를 얻었고, 두 번째 경우에서는 동일한 가중치를 갖는 키팩트로 검색한 결과의 정확도와 다른 가중치를 갖는 키팩트로 검색한 결과의 정확도가 같게 나왔다 그렇지만, 이 정확도는 단순히 15위 내에 포함되어 있는지를 조사한 결과다. 검색된 결과의 순위를 하나하나 살펴보면 다른 가중치를 갖는 키팩트로 검색한 결과의 순위가 미리 만들어 놓은 질의 응답의 순위에 가장 근접했다 그리고, 미리 정의해 놓은 40개의 질의 응답에 대한 평균 정확도는 다음과 같다

	키워드	동일한 가중치를 갖는 키팩트	다른 가중치를 갖는 키팩트
정확도	69%	74%	78%

6. 결론 및 향후 연구방향

다중단어인 키팩트를 이용하여 정보를 검색할 경우에 모든 키팩트에 같은 가중치를 부여해서 검색을 하는 것보다는 각 키팩트에 서로 다른 가중치를 부여해서 검색하는 것이 정확도에 있어서 더 좋은 결과를 얻을 수 있었다 키팩트에 각각 다른 가중치를 부여한 것이나 가중치를 동일하게 부여한 것이나 정확도에 있어서는 그다지 큰 차이를 보이지는 않았지만, 검색된 결과의 순위에 있어서는 각각 다른 가중치를 부여한 키팩트를 사용한 경우에 더 적합한 문서를 검색할 수 있었다 그리고 향후에 연구할 방향으로는 가중치를 두 가지 경우로만 생각할 것이 아니라, 각각의 경우에 맞게 가중치를 부여하는 방법이나 알고리즘이 개발되어야 한다.

참고 문헌

[1] C.J. van Rijsbergen *Information Retrieval*, Butterworths, London, Second Edition, 1979

[2] 정경덕, 최동시, 전미선, 서래원, 박세영, "의미기반 정보검색을 위한 ETRI-NLPS 자연어처리 형태 태그 세트," 한국전자통신연구원 자연어처리연구실 1997

[3] G Salton and M. G McGill, *Introduction to Modern Information Retrieval*, McGrawHill, New York, 1983

[4] "내용기반 멀티미디어 정보검색 기술 개발", 한국전자통신연구원.

[5] 장대석, "키팩트개념기반 정보검색시스템", 한양대 전자계산학과 석사학위 논문, 1997