개념 공간을 이용한 의미 인덱싱

강보영* 김혜정 황선욱 이상조
경북대학교 컴퓨터 공학과
comeng99@hotmail.com

Semantic Indexing Using Concept Space

Bo-Yeong Kang* Hae-Jung Kim Sun-Wook Hwang and Sang-Jo Lee
Department of Computer Engineering, Kyungpook National University

요 약

본 논문은 문서내의 의미적인 관계에 기반하여, 문서의 내용을 보다 잘 추측할 수 있는 의미 인덱스 추출 및 가중치 부여 시스템을 제안하고자 한다. 문서 내의 개념 추출에 있어서는 기존의 어휘 체인(lexical chains)에 관한 연구를 확장하여 적용였다. 또한, 추출된 개념에서 중요 어휘에 가중치를 부여하기 위해서, 개념 벡터 공간을 이용한 정보성(information quantity)과 정보비(information ratio)를 정의하고, 인덱스의 가중치를 측정할 수 있는 정량화 할수 있는 척도로 제시하였다.

1. 서론

인터넷 상에서 정보 검색 시스템은 사용자가 필요로하는 정보를 찾기 위하여 문서의 내용 뿐만 아니라 사용자의 의도를 파악하고, 주어진 질의에 적합한 문서를 검색해야한다. 그러나 인터넷 상의 방대한 문서들은 문서를 대표하는 색인어가 없는 경우가 많고, 대표 색인어를 가진 문서라도 그 문서의 의미적인 내용을 표현하는 좋은 지시자나 색인어가 있다면, 텍스트 마이닝, 문서요약, 문서 클러스터링과 같은 다양한 연구 분야에 많은 도움을 줄 것으로 판단된다. 본 논문에서는 문서내의 의미적인 관계에 기반하여 문서의 내용을 추측하고 보다 정확한 정보를 찾는데 도움을 줄 수 있는 효과적인 색인어 추출 및 가중치 부여 시스템을 제안하고자 한다.

본 논문은 문서를 여러 개념들이 복합적으로 이루어낸 하나의 복합적인 개념(Concept)으로 간주한다. 따라서, n개의 개념으로 이루어진 개념 벡터 공간(Concept Vector Space)을 정의하고, 문서는 n차원의 개념 벡터 공간상에 존재하는 점으로 인식한다. 개념 벡터 공간을 구성하는 개념의 추출에 있어서는 기존의 어휘 체인(lexical chains)에 관한 연구를 확장하여 사용한다. 추출된 개념으로 이루어진 개념 벡터 및 문서를 표현하는 문서벡터를 사용하여 의미 색인어(semantic index)를 추출하고, 추출된 색인어가 문서 내에서 가지는 의미적인 중요도에 대한 정도를 반영할 수 있는 의미 색인어 가중치(Semantic index weight)를 부여한다. 제안된 방법은 추출된 키워드가 문서 내에서 가지는 의미적인 중요도에 대한 정도를 반영할 수 있으므로, 문서 내에서 단어의 중요도를 통계적인 방법만으로 해결하였던 용어빈도(term frequency)기반 가중치 기법을 대안 할 수 있을 것으로 기대된다.

2. 관련연구

색인 및 가중치 부여에 관한 선행 연구들의 대부분은 통계적인 방법에만 의존하고 있으며, 이러한 접근법이 정확한 색인어를 추출하는데 한계를 주고 있다[1]. 용어 빈도(term frequency)는 긴 문서에서는 좋은 방법이 되지만, 짧은 문서에서는 성능이 좋지 않을 뿐 아니라, 대용어나, 동의어 등을 찾지 못하므로 용어 빈도를 정확하게 표현하지 못한다. 역문헌 빈도(inverse document frequency)는 참조하는 문서 집합(collection)이 계속해서 변화하면, 색인어의 가중치를 매번 재계산해야 하는 불편이 있다. 길이 정규화(length normalization)은 각 문서가 다른 길이를 가지기 때문에 긴 문서에서 용어 빈도가 큰 값을 가지게 되는 문제를 감소시키기 위해 제안된 방법이다. 그러나 이 접근법 또한 용어빈도를 기본적으로 사용하고 있으므로, 용어빈도가 가지는 문제점을 가질 수 밖에 없다.

또한 어휘 체인 관련 연구에서, Barzilay and Elhadad는 어휘 체인을 문서요약에 적용하는데 있어, 대표단어(representative words)를 선별하는 방법을 제안하였다[2]. 이것은 문서에서 화제를 추출하는 것으로, 어휘 체인을 이용하여 색인어를 추출할 수 있는 접근법으로도 간주될 수 있다. 대표 단어(representative words)란 그 체인을 대표하는 체인 멤버로서 강한 체인에 나타나는 단어들 중 평균 빈도(average frequency) 이상 발생한 단어를 말한다. Barzilay and Elhadad의 이러한 접근법은 문서 요약에 적용할 중요 명사를 탐색하는 과정에서 취했던 의미론적인 접근에서 벗어나 평균 빈도라는 통계적인 접근을 모색하였다. 또한 대표 단어들이 문장에서 차지하는 의미적인 중요도가 다름에도 불구하고 그러한 차이점을 전혀 반영하지 않았다. 즉, 점수가 높은 체인 내의 대표 단어들은 점수가 낮은 체인내의 대표 단어들보다 문서의 의미적인 내용을 표현할 화제(topic)일 가능성이 크다. 그럼에도 불구하고 단순히 빈도만 고려함으로써 점수가 높은 체인에서 선별된 단어와 점수가 낮은 체인에서 선별된 단어의 차이를 밝혀 낼 수 없었다. 예를 들면, 체인 점수가 20인 체인에서 10번 나타난 단어와, 체인 점수가 10인 체인에서 10번 나타난 단어는 문서 내에서 차지하는 의미적인 중요도가 다르지만, 차이점을 전혀 반영하지 못한다.

본 논문은 색인어를 추출하고 가중치를 부여하는데 개념 벡터 공간을 이용한 의미적인 접근을 제안함으로써 기존의 방법론들이 가졌던 한계를 해결하고자 하였다. 먼저, 제안된 방법은 색인어 추출에 있어서 어휘 체인을 이용하여 연관된 어휘들을 고려함으로써 용어 빈도(term frequency)가 가지는 문제점을 해결할 수 있다. 또한, 전체 문장이 가지는 정보를 1로 보고 그것에 대해 단어가 가지는 의미적 정보의 비율로 가중치를 표현함으로써, 문장 길이에 독립적으로 가중치를 부여할 수 있는 장점이 있다.

3. 개념 공간을 이용한 의미 인덱스 추출

제안된 시스템은 다음과 같이 동작한다: 먼저 색인어를 추출하고자 하는 문장으로부터 명사를 추출한다. 추출된 명사는 다양한 의미를 갖고 있으므로, 베이지안 분류(Bayesian classification)을 이용하여 명사의 의미를 결정한다. 그런 후, 의미가 결정된 명사들로 어휘 체인을 생성하여 어휘 체인 내에 있는 각 명사들과 체인들에 점수를 부여한다. 이때, 많은 어휘 체인들 중 문서를 대표하는 대표 체인(representative chains)을 찾아낸다. 추출된 대표 체인, 즉, 대표 개념들로 전체 문서 개념 벡터를 구성하고, 문서 벡터를 기준으로 각 개념 및 개념 내의 어휘들이 가지는 중요도를 계산한 후, 의미 색인어를 추출하여 문서 내에서의 중요도를 반영하는 가중치를 부여한다.

3.1 어휘 체인과 개념 공간

어휘 체인(lexical chains)이란 문서에서 관련된 단어들을 연결한 체인으로, 문서 내 응집성의 구조를 밝힐 수 있다. 본 논문은 어휘 체인이 담화 구조를 표현할 수 있다는 언어학적인 연구로부터[3][4], 연관된 단어들의 연결인 어휘 체인이 문서를 표현하는 하나의 개념으로 사용된다고 가정한다. 따라서, 문서를 구성하는 개념들은 어휘 체인에 의해 추출될 수 있다. 그러나, 문제를 단순화 시키기 위하여, 모든 어휘체인을 문서를 대표하는 개념으로 간주하지 않고 대표 체인을 선별한다. 대표 체인(Representative Chain)이란 문서로부터 구성되는 여러 체인들 중 문서의 개념을 대표할 수 있는 체인이다. 세 개의 대표 체인으로 이루어진 예제 문서를 살펴보자. 문서에서 대표 체인으로 선별된 체인이 체인1, 체인2, 체인3이라고 가정한다. 각 체인은 사람, 마취, 기계를 나타내는 서로 다른 개념(concept)을 표현하고 있다. 따라서 문서는 강한 체인1, 2, 3이 나타내는 개념1, 개념2, 개념 3으로 이루어져 있고, 각 개념은 문서내의 어휘들에 의해서 이루어진다. 각 개념에 속한 모든 어휘들이 색인어로 사용되는 것이 아니라, 문서 및 개념들의 중요도를 고려하여 색인어를 추출한다.

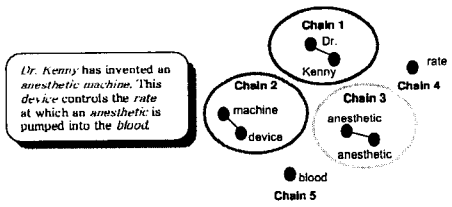

Fig. 1. 예제 문서의 어휘 체인

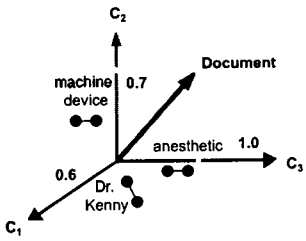

Fig. 2. 그림 1의 개념 벡터 공간 표현

그림 1을 개념 벡터 공간에 표현하면 그림 2와 같다. 개념 벡터 공간은 문서의 개념을 표현할 수 있는 독립적인 개념축(Independent concept axis)들로 이루어져있고, 각 개념 축에 표현되는 개념 벡터의 크기에 따라 전체 문서의 개념 벡터의 방향 및 크기가 결정된다.

3.2 어휘 체인을 이용한 개념 추출

문서를 구성하는 개념을 추출하기 위하여 문서 내 명사들로 어휘 체인을 생성하고, 각 체인 및 어휘에 점수를 부여한다. 점수가 부여된 어휘 체인 중 문서의 개념을 대표하는 대표 체인을 선별한다. 어휘 체인 내의 각 체인과 명사에 점수를 부여하는 방법은 다음과 같다. 이 때, 체인 내에서 명사가 가지는 관계는 반복, 동의어, 상위어, 하위어, 부분어 등의 네 가지 유형이다.

체인내의 명사 N_i의 점수$S_{NOUN}(N_i)$는 수식과 같다. $NR^k_{N_i}$는 명사 N_i가 가지는 관계k의 갯수를 의미한다. $SR^k_{N_i}$은 관계 k의 가중치를 표현한다.

$$S_{NOUN}(N_i) = \sum_k (NR^k_{N_i} \times SR^k_{N_i}) \quad (1)$$

체인 Ch_x의 점수 $S_{CHAIN}(Ch_x)$는 다음과같다:

$$S_{CHAIN}(Ch_x) = \sum_{i=1}^{n} S_{NOUN}(N_i) + penalty \quad (2)$$

대표 체인 기준은 다음과 같이 정의된다:

$$S_{CHAIN}(Ch_i) \geq \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} S_{CHAIN}(Ch_i) \quad (3)$$

3.3 정보성과 정보비를 이용한 의미 인덱스

본 절에서는 대표 개념으로부터 색인어를 선별하고 의미적 가중치를 부여할 수 있도록, 개념 벡터 공간을 사용하여 전체 문서를 이루는 개념 및 단어의 의미적인 중요도를 정량화 하고자 한다. 각 개념들의 의미적인 중요도는 벡터 공간의 특질을 이용한 수식을 기반으로 유도될 수 있다. 개념 벡터 C_1의 크기가 a이고, C_2의 크기가 b일 때, 두 개념의 조합으로 유도된 문서 전체의 개념을 나타내는 벡터의 크기는 $\sqrt{a^2+b^2}$이다. 문서 개념 벡터에서, 개념 벡터 C_1이 문서 벡터를 구성하는 크기는 x이고, 개념벡터 C_2가 문서벡터를 구성하는 크기는 y이다.

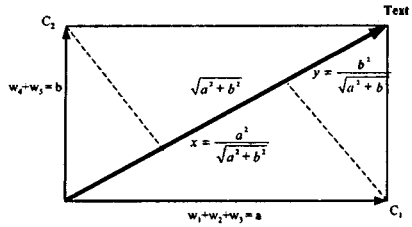

Fig. 3. 벡터 공간 특성

벡터 공간의 특성을 이용하여 단어 정보성과 단어 정보비가 유도된다. 단어 정보성(Ω_{W_j})은 단어가 문서에 대해서 가지는 의미적인 정보의 양이다. 단어 정보비($\Psi_{W_j|T}$)는 문서가 가지는 정보성에 대하여, 비교 대상이 되는 단어가 가지는 정보성의 비율이다.

$$\Omega_{W_j} = \frac{W_j \cdot C_i}{\sqrt{\sum_k C_k^2}} \quad (4)$$

$$\Psi_{W_j|T} = \frac{W_j \cdot C_i}{\sum_k C_k^2} \quad (5)$$

추출된 대표 개념으로부터 전체 문서 벡터에 대한 각 개념과 단어의 정보성 및 정보비를 계산한 후, 의미 색인어를 추출하고 의미 색인어 가중치를 부여한다. 의미 색인어 추출 기준은 다음과 같다. 추출된 의미 색인어가 가지는 가중치는 해당 색인어가 가지는 정보비($\Psi_{W_j|T}$)이다.

$$\Omega_{W_j} \geq M + C \cdot \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\Omega_{W_i} - M)^2} \quad (6)$$

where, $M = \frac{1}{m}\sum_{i=1}^{m}(\Omega_{W_i})$

4. 실험 및 분석

제안된 의미 인덱스의 신뢰성과 효율성을 보이기 위해서, 다섯가지의 문서 집합에 대해서 기존의 인덱스들과 성능 비교 실험을 수행하였다. 먼저 여섯 명의 피실험자에게 다섯 문서로부터 인덱스를 추출하고 각 인덱스에 가중치를 부여하도록 하였다. 다섯 문서는 각각 특별한 주제에 대하여 논하고 있었지만 대부분 "운동"이라는 일반적인 주제를 내포하고 있었다. 표 I로부터 피실험자들이 "운동"이라는 인덱스에 대하여 부여한 가중치와 제안된 접근이 부여한 의미 가중치(semantic weight, SW)가 term frequency(TF), length normalization(LN)에 비해 좋은 결과를 보임을 알 수 있다. 그림 4는 표 I를 챠트로 표현한 것이다. 그림 4는 standard TF와 제안한 가중치 기법이 피실험자들이 매긴 가중치 그래프와 매우 유사한 성능을 보임을 나타낸다 . 그러나 standard TF는 제안한 기법에 비해 문서들 사이의 미묘한 의미적인 차이를 구별하지 못한다. 즉, 피실험자들은 문서1에서 "운동"이 가지는 의미적인 중요도가 문서 2에서 가지는 중요도보다 높다고 평가하였으나, standard TF는 이러한 점을 반영하지 못하고 있다.

TABLE I
인덱스 *exercise*에 대한 가중치 비교

Text	User weight	TF	LN	Standard TF	SW
1	0.39	3	0.428	0.29	0.3748
2	0.31	3	0.75	0.375	0.2401
3	0.25	1	0.33	0.18	0.1320
4	0.11	1	0.125	0.11	0
5	0	1	0.2	0.12	0

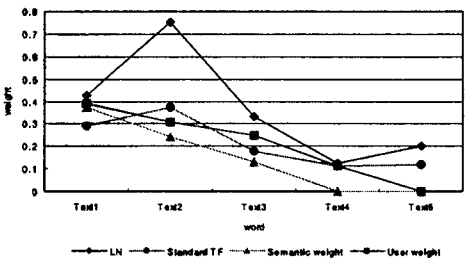

Fig. 4. 표1에 대한 가중치 비교

5. 결론

본 논문은 문서를 구성하는 개념에 기반하여, 문서의 의미적인 내용을 보다 잘 표현할 수 있는 의미 인덱싱 기법을 제안하였다. 문서 내의 개념 추출에 있어서는 기존의 어휘 체인(lexical chains)에 관한 연구를 확장하여 적용였다. 또한, 인덱스가 문서 내에서 가지는 의미적인 중요도를 반영하는 가중치를 측정할 수 있는 척도로, 개념 벡터 공간을 이용한 정보성(informativeness)과 정보비(information ratio)를 정의하였다. 제안된 방법은 추출된 키워드가 문서 내에서 가지는 의미적인 중요도에 대한 정도를 반영할 수 있으므로, 단어의 중요도를 통계적인 방법만으로 해결하였던 용어빈도(term frequency)기반 가중치 기법을 대안 할 수 있을 것으로 기대된다.

References

[1] M.-F. Moens, Automatic Indexing and Abstracting of Document Texts, Kluwer Academic Publishers, 2000.

[2] R. Barzilay, M. Elhadad, Using lexical chains for text summarization, Proc. ACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

[3] J. Morris, Lexical cohesion, the thesaurus, and the structure of text, Master's thesis, Department of Computer Science, University of Toronto, 1988.

[4] J. Morris, G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17(1)(1991) 21-43.