논 문
58-3-30

가중치 기반 PLSA를 이용한 문서 평가 분석

Reputation Analysis of Document Using Probabilistic Latent Semantic Analysis Based on Weighting Distinctions

조 시 원* · 이 동 욱†
(Shiwon Cho · Dong-Wook Lee)

Abstract - Probabilistic Latent Semantic Analysis has many applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. In this paper, we propose an algorithm using *weighted Probabilistic Latent Semantic Analysis Model to find the contextual phrases and opinions from documents*. The traditional keyword search is unable to find the semantic relations of phrases, Overcoming these obstacles requires the development of techniques for automatically classifying semantic relations of phrases. Through experiments, we show that the proposed algorithm works well to discover semantic relations of phrases and presents the semantic relations of phrases to the vector-space model. The proposed algorithm is able to perform a variety of analyses, including such as document classification, online reputation, and collaborative recommendation.

Key Words : PLSA, Reputation analysis, EM-Algorithm

1. 서 론

인터넷의 발달로 사용자들은 온라인 커뮤니티 서비스나 블로그를 이용하여 자신의 의견을 적극적으로 표현하고, 각종 검색 서비스를 이용하여 다른 사용자들의 의견을 검색함으로써 서로의 의견을 공유한다. 이러한 사용자들의 적극적인 참여는 그 정보에 대한 신뢰를 형성하는 중요한 요인이 되기 때문에, 온라인에서 생산되는 정보와 그 정보에 대한 다양한 평가 의견은 대단히 중요하다. 평가 분석 (reputation analysis)은 평가 의견을 포함하는 문서에 나타난 다양한 문장들을 이용하여, 문장 내에서 의견을 나타내는 단어를 분석함으로써 의견에 대한 전반적인 평가를 하고, 문서에 나타난 주제어와 의견을 바탕으로 주제어와 관심 표현이 비슷한 패턴을 가진 문서들을 분류 (classification)하는 방법이다. 이러한 평가 분석을 정보 검색에 적용하면 사용자가 질의한 정보에 대하여 여러 사용자들의 평가를 검색함으로써, 사용자의 다양한 욕구를 충족시켜줄 수 있게 된다.

평가 분석은 주어진 문서를 형태소 분석과 구문 분석을 통하여 문서가 담고 있는 주제어와 평가 표현을 의미하는 서술어를 추출한 다음, 분석 알고리즘을 통하여 주제어와 평가 의견(서술어)의 관계를 분석하여 전반적인 평가를 도출한다. 본 논문에서는 문서에서 추출한 명사와 서술어의 출현 빈도를 이용한 통계적 기법이 아닌 문서에 출현하는 단어의 특징적인 패턴을 이용한 잠재토픽모델 (latent topic model)을 분석 알고리즘으로 사용하였다[1].

잠재토픽모델 (latent topic model)은 공기 (co-occurrence) 정보를 이용하여 단어의 특징적인 패턴과 단어 사이의 의미 있는 상관 관계를 분석하는 방법이다. 잠재토픽모델은 주어진 데이터로부터 중요 특성을 분석하여 요약적으로 제시하며, 분류 작업에서도 효과적으로 적용이 가능하다. 이와 같은 특성을 이용하여 잠재토픽모델은 대용량의 데이터를 분석하기 위한 특징점의 자동 추출 및 군집화/가시화 등의 응용에 널리 활용되었으며, 최근에는 정보 검색, 이미지/음성 인식, 생물정보학 (bioinformatics), 멀티미디어 데이터 분석 등의 분야에 많이 사용되고 있다. 대표적으로 PLSA (Probabilistic Latent Semantic Analysis)[2], LSA (Latent Semantic Analysis)[3], NMF(Non-negative Matrix Factorization)[4] 등이 있다. PLSA는 공기 데이터 분석을 위한 통계적 기법으로 데이터에 대한 생성 모델을 다항분포로 정의하고 EM(Expectation Maximizatioin) 알고리즘을 이용하여 유사도 값을 최대화하도록 모델을 학습한다. LSA는 단어의 출현 빈도에 따라 문맥 정보에 기반하여 잠재의미를 분석하는 방법으로, 학습모델은 SVD (Singular Value Decomposition)를 이용한다. NMF는 인간이 객체의 부분 정보의 조합으로 객체의 전체를 인식하는 것에 착안하여, 각 객체를 구성하고 있는 부분 정보를 사전 정보 없이 학습하는 부분 기반 비감독 학습 방법이다.

본 논문의 목적은 문서를 단어의 집합으로 보는 기존의 키워드 검색의 관점에서 벗어나, PLSA모델을 이용하여 주어와 목적어, 서술어 등과 같이 구문적으로 연결된 각 단어들의 연관 관계를 이용하여 단어의 의미와 문서의 내용을 통계적으로 분석하기 위한 것이다. 본 논문의 구성은 다음

† 교신저자, 정회원 : 동국대 공대 전기공학과 교수 · 공박
E-mail : dlee@dongguk.edu
* 정 회 원 : 동국대 공대 전기공학과 박사과정
접수일자 : 2008년 11월 4일
최종완료 : 2009년 2월 5일

과 같다. 2장에서는 PLSA를 이용한 평가 분석 방법에 대하여 설명한다. 3장에서는 제안된 알고리즘을 이용한 실험 및 결과에 대해 기술하고, 마지막으로 4장에서는 결론 및 향후 연구 방향을 제시한다.

2. 본 론

2.1. 구문 분석

문서들을 문장 단위로 분리한 후 형태소 해석기와 구문 분석기[5]를 이용하여, 문장에 등장하는 단어들의 구문적 관계와 자질 정보를 분석한다. 각 단어 w는 구문 분석을 통하여 자질 정보(주격, 목적격 , 관형어, 보어, ...)와 품사 정보(명사, 동사, 형용사, ...)를 갖게 된다. 구문 분석을 통해 두 단어 사이의 구문적 연결 관계를 분석하여 다음 식(1)과 같이 의미 연결(semantic link) 관계를 생성한다.

$$\begin{aligned} SL(w_i,w_j) = & < LCD \rightarrow 모니터 >, \\ & < 저렴한것이 \rightarrow 장점이다 >, \\ & < DSLR \rightarrow 렌즈 >, \\ & < DSLR \rightarrow 화질 > \end{aligned} \tag{1}$$

2.2. PLSA (Probabilistic Latent Semantic Analysis)

Probabilistic LSA는 공기(co-occurrence) 데이터 분석을 위한 통계적 기법으로 정보 검색, 정보 분류 등의 분야에서 많이 응용된다. N개의 문서가 있고, 각 문서에 M개의 단어가 존재한다면, 문서 집합 $D=\{d_1,d_2,...,d_N\}$와 단어 집합 $W=\{w_1,w_2,...,w_M\}$으로 표현할 수 있다. 문서와 단어 집합은 $N\times M$ 문서-단어 행렬 (document-term matrix) $DW=|n(d_i,w_j)|_{NM}$로 표현할 수 있다. 여기서, $n(d_i,w_j)$는 문서d_i에서 단어w_j의 출현 회수(term frequency)를 의미한다. PLSA는 각각의 문서d와 단어w의 집합과 연관되는 잠재 의미 집합 $z\in Z=\{z_1,z_2,...z_k\}$을 이용하는 모델이다[그림 1]. z는 문서에 등장하는 각 단어에 대응하는 의미 토픽(semantic topic)을 의미한다.

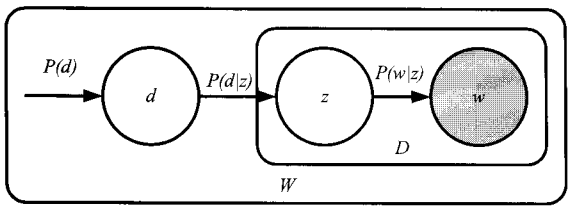

그림 1 PLSA 모델
Fig. 1 PLSA model.

PLSA모델은 다음과 같은 과정을 거친다.

$P(d_i)$를 이용하여 문서 집합 D에서 문서 d_i를 선택한다.

$P(z_n|d_i)$를 이용하여 잠재 집합 z_n을 선택한다.

$P(w_j|z_n)$를 이용하여 단어 집합 W에서 z_n에 속하는 w_j를 구한다.

이러한 과정의 결과로 z_n에 대응하는 문서d_i와 단어w_j의 문서-단어 쌍 (d_i,w_j)를 얻게 되며, 이것을 확률 모델로 표현하면 다음과 같다.

$$\begin{aligned} P(d_i,w_j) &= P(d_i)P(w_j|d_i), \\ P(w_j|d_i) &= \sum_{z\in Z} P(w_j|z)P(z|d_i) \\ P(d_i,w_j) &= \sum_{z\in Z} P(z)P(d_i|z)P(w_j|z) \end{aligned} \tag{2}$$

$P(w_j|z_n)$와 $P(z_n)$를 이용하여 문서d_i가 잠재 의미 집합(카테고리) z_n에 속할 확률 $P(z_n|d_i)$를 구한다. $P(z_n|d_i)$는 식 (3)와 같이 베이지안 정리를 이용하여 구할 수 있다.

$$\begin{aligned} P(z_n|d_i) &= \frac{P(d_i|z_n)P(z_n)}{P(d_i)} \\ &= \frac{P(d_i|z_n)P(z_n)}{\sum_{z\in Z} P(d_i|z)} \end{aligned} \tag{3}$$

$P(w_j|z)$는 의미 토픽에 속한 단어의 분포를 나타내며, $P(d_i|z)$는 의미 토픽에 대한 문서의 분포를 나타난다. z가 주어질 때, w_j와 d_i는 조건부 독립 조건을 만족한다. PLSA 학습 모델은 유사도 값이 최대가 되도록 하는 $P(z_n)$, $P(w_j|z_n)$, $P(z_n|d_i)$를 구하는 것이다. 유사도(likelihood)함수 $L(D,W)$은 다음과 같다.

$$\begin{aligned} L(D,W) &= \log \prod_{d\in D} \prod_{w\in W} P(w|d)^{n(d,w)} \\ &= \sum_{d\in D} \sum_{w\in W} n(d,w) \log P(d,w) \end{aligned} \tag{4}$$

EM(Expectation Maximizatioin) 알고리즘은 잠재 토픽 모델에서 많이 알려진 최대 유사도 추정 방법이다[6]. PLSA 학습모델에서 사용하는 EM-알고리즘의 E(Expectation) 단계와 M(maximization) 단계는 다음과 같이 정의된다 [7]-[11].

$$\begin{aligned} E-step: P(z_n|d_i,w_j) &= \frac{P(z_n)P(d_i|z_n)P(w_j|z_n)}{\sum_{z\in Z} P(z)P(d_i|z)P(w_j|z)} \\ &= \frac{P(d_i|z_n)P(w_j|z_n)}{\sum_{n\in Z} P(d_i|z)P(w_j|z)} \end{aligned} \tag{5}$$

$$\begin{aligned} M-Step: P(w_j|z_n) &\propto \sum_{d\in D} n(d,w_j)P(z_n|d,w_j) \\ P(z_n|d_i) &\propto \sum_{w\in W} n(d_i,w)P(z_n|d_i,w) \\ P(z_n) &\propto \sum_{d\in D} \sum_{w\in W} n(d,w)P(z_n|d,w) \end{aligned} \tag{6}$$

EM알고리즘은 식 (5)와 식 (6)을 반복적으로 실행하여 지역 최적해를 추정한다[그림 2].

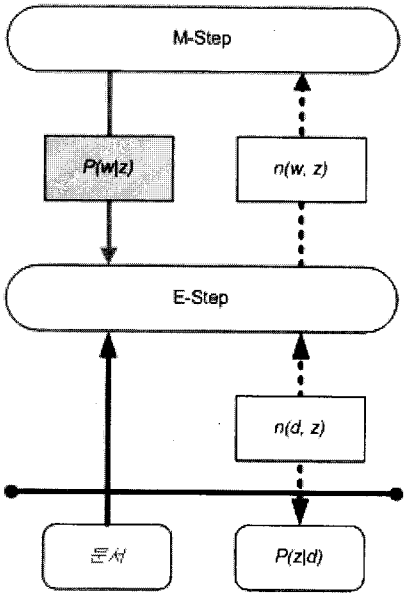

그림 2 EM알고리즘에 의한 PLSA 학습
Fig. 2 PLSA learning with EM algorithm.

2.2. 가중치 적용

많은 문서들 중에서 그 문서를 대표하는 특징을 추출하기 위해 단어의 출현 빈도가 많이 이용된다. 그러나, 실제로 많은 문서에서 그 문서를 대표하는 단어의 출현 빈도는 높지 않다. 상대적으로 출현 빈도가 높은 단어는 일반적인 의미로 사용되므로 문서를 대표하는 단어로 볼 수 없다. 문제점을 해결하기 위하여 문서-단어 행렬에서 단어의 출현 회수 $n(d_i,w_j)$에 가중치를 적용하여 PLSA학습을 수행한다. 단어의 출현 회수 $n(d_i,w_j)$에 가중치를 적용한 $n(d_i,w_j)'$ 함수를 다음과 같이 정의한다[12]-[14].

$$n(d_i,w_j)' = \frac{n(d_i,w_j)}{\sqrt{\sum_{d\in D}(n(d,w_j))^2}} log\frac{N}{df(w_j)} \quad (7)$$

$n(d_i,w_j)$: 문서 d_i에 속한단어 w_j의출현회수
N: 전체문서수
$df(w_j)$: 단어 w_j가들어있는문서의수

PLSA의 학습 결과로부터 생성된 잠재 의미 집합 z는 문서 분류를 위한 카테고리를 의미한다. 카테고리에 속한 단어의 분포 $P(w_j|z_n)$는 해당 카테고리의 특징을 나타내는 특징점에 해당한다. 그러나, 출현 빈도가 높고, 전체 카테고리에 공통적으로 등장하는 단어일수록 카테고리의 특징을 나타내는 특징점으로 볼 수 없다. 특히 단어의 분포가 서로 유사한 카테고리의 경우, 카테고리를 구분할 수 있는 변별력은 낮아진다. 그러므로, 카테고리 사이의 변별력을 높여주기 위해 단어의 출현 회수에 다음과 같이 가중치를 적용한다.

$$n(z_n,w_j)' = \frac{n(z_n,w_j)}{\sqrt{\sum_{z\in Z}(n(z,w_j))^2}} log\frac{N}{cf(w_j)} \quad (8)$$

$n(z_n,w_j)$: 카테고리 z_n에 속한단어 w_j의출현회수
N: 전체카테고리수
$cf(w_j)$: 단어 w_j가 들어있는카테고리수

가중치가 적용된 $n(z_n,w_j)'$을 사용하여 새로운 $P(w_j|z_n)$를 식 (9)와 같이 구하면, 카테고리의 특징점에 해당하는 단어에 대한 변별력을 높이고, 출현 빈도는 높지만 전체 카테고리에 공통적으로 출현하는 단어에 대한 변별력을 낮추게 된다. 이런 특징을 이용하여 단어의 분포가 유사한 카테고리들 사이의 변별력을 높일 수 있다.

$$P(w_j|z_n)' = P(w_j|z_n)\times n(z_n,w_j)' \quad (9)$$

2.3. 단어의 연관 관계 계산

문서 분류 (document classification)에서는 서로 공기하는 명사 단어의 연관 관계만 고려하지만, 평가 분석에서는 '주어+목적어','주어+서술어,'목적어+서술어'등과 같이 구문적으로 단어들의 의미 관계를 분석해야 한다. 단어의 의미 관계는 크게 세 가지 형태로 구분할 수 있다.

주제어(noun): LCD 모니터'와 같은 상품 분류나 '가격', '성능', '크기', '디자인' 등과 같이 평가대상의 특징을 의미하는 단어들이 주제어에 해당된다.

서술어(predicate) : 주제어에 대한 의미 표현을 기술하는 단어들이다. '크다', '작다', '빠르다'등과 같이 주제어에 대한 특징과 속성을 표현한다. '작지 않다', '좋지 않다'와 같이 의미 표현이 반전되는 경우, 의미 표현 정도를 반대로 바꾸어 주어야 한다. 이 경우, 보조 용언도 같이 포함하여 평가 표현에 대한 긍/부정을 판단한다.

가중 표현: 주제어에 대한 서술 정보와 관계없이 언어적 표현에서 부가적인 의미를 부여하는 단어들이다. '조금', '대단히', 'ㄹ것 같다'등과 같이 의미 표현의 정도에 영향을 주는 단어들이다.

주제어, 서술어, 가중 표현의 관계는 그림 3과 같이 표현할 수 있다. '주제어+주제어'의 경우, 카테고리 내에서 서로 연관 단어로 볼 수 있으며, '주제어+서술어'는 주제어에 주로 사용되는 의미 표현 단어를 분석할 수 있다. 가중 표현은 서술어의 의미 표현에 영향을 준다.

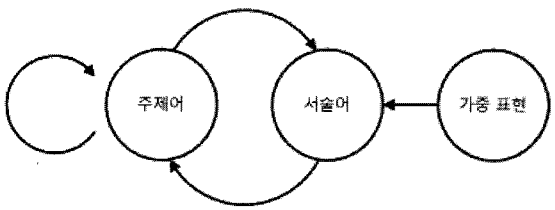

그림 3 주제어, 서술어, 가중 표현의 상관 관계도
Fig. 3 Correlation diagram of topic, predicate, and weighted representation.

$P(w|z)$에 의해 카테고리에 속한 두 단어 (w_i, w_j)의 연관 관계는 식 (10)을 이용하여 계산한다. (w_i, w_j)가 주제어가 아닌 서술어나 가중 평가 단어의 연관 관계 식(1)의 의미 연결 관계를 이용하여 선택적으로 연관 관계를 계산한다. 연관 관계를 적용하는 조건은 표 1과 같다.

$$Sim(w_i, w_j) = \sum_{z \in Z} P(w_i|z)P(w_j|z)$$
$$p(w_i|z_n) = \frac{P(w_i)P(z_n|w_i)}{\sum_{z \in Z} P(w_i)P(w_i|z)} \quad (10)$$

표 1 연관 관계 적용 조건

Table 1 Applied conditions to the words of association.

(주제어, 주제어)	식 (10)적용
(주제어, 서술어) (서술어, 주제어)	의미 연결 관계가 있는 경우, 식 (10)적용
(가중 표현, 서술어)	의미 표현 강도 조정

연관 관계를 계산하여, 각 카테고리 별로 연관 단어 집합 C_n을 생성할 수 있다. 표 2는 생성된 연관 단어 집합의 결과이다.

$$C_n = \{(w_i, w_j, wc, v, r) | w_i \in z_n, w_j \in z_n, wc = \mathrm{N}, v = \mathrm{R}, r = \mathrm{R}\}$$
wc: 단어쌍의 출현횟수
v: 연관정도
r: 평가도, 초기값 : 0 (11)

표 2 카테고리별 연관 단어 집합

Table 2 Associative word set by categories.

카테고리	연관 단어 집합: {주제어, 주제어},{주제어, 서술어}
z_1	{LCD, 모니터}, {모니터, 해상도}, {해상도, 높다}, {가격, 비싸다}, {화질, 만족한다}, {화질, 어둡다}, {화질, 밝다}, {가격, 착하다}, {저렴하다, 가격}, {가격, 저렴하다}, {가격, 저렴하지 않다}, {가격, 저렴하다}, …
z_2	{DSLR, 카메라}, {이미지, 화질}, {기능, 장점이다}, {가격, 비싸다}, {성능, 만족하다} {기능, 부족하다}, …

2.4. 평가 표현 추출

연관 단어 집합 C_n에 속한 주제어와 평가 단어를 군집화한 다음, C_n의 평가도 r을 다음과 같이 계산한다.

$$r = \begin{cases} \dfrac{w_i\text{의출현횟수}}{\text{전체 의미 표현 단어의 수}}, & \text{if } w_i = \{\text{서술어, 부가표현}\} \\ 0, & \text{if } w_i = \{\text{주제어}\} \end{cases} \quad (12)$$

평가도 r을 계산한 다음, 평가도 순으로 군집화하면 표 3과 같은 결과를 얻게 되며, 주로 사용되는 표현의 대상이 되는 주제어와 표현 관계를 분석 할 수 있다. 또한 평가도 r을 이용하여, 주제어와 평가 단어 사이의 관계를 벡터 공간 모델로 표현이 가능하다.

표 3 연관 단어 집합 C_n의 평가도 계산

Table 3 Rating of associative word set C_n.

카테고리	단어 쌍	평가도
z_1	{LCD, 모니터}	0
	{모니터, 해상도}	0
	{해상도, 높다}	0.8
	{화질, 어둡다}	0.1
	{화질, 만족한다}	0.5
	{화질, 밝다}	0.4
	{가격, 비싸다}	0.68
	{가격, 저렴하다}	0.26
	{가격, 착하다}	0.06
	…	…
z_2	{DSLR, 카메라}	0
	{이미지, 화질}	0
	{기능, 장점이다}	0.3
	{가격, 저렴한}	0.6
	{성능, 만족하다}	0.6
	{기능, 부족하다}	0.7
	…	…

'주제어+주제어'과 같은 단어 쌍에 대한 평가도는 동등 관계이므로, 0으로 설정하였다. 표 3에서 클래스 C1의 평가 주제를 {LCD, 모니터}로 선정한다면, 화질은 만족하지만, 가격은 비싸다는 평가 표현을 측정할 수 있다.

표 4와 같이 평가 표현에 대한 평가 단어 사전을 이용하면 주제어에 대한 평가도(긍정, 중립, 부정)를 측정할 수 있다.

표 4 평가 표현 사전

Table 4 Dictionary of reputation words.

부정	중립	긍정
-1 … …	0	… … … +1
비싸다		저렴하다, 싸다
장점		단점
어둡다		밝다
차다		뜨겁다
낮다		높다
여성		남성
낡다		새롭다

평가 단어 사전은 그림 4와 같이 문장의 서술어를 분석하여 평가 단어를 수집하고, 평가 단어와 평가 표현 정도(긍정, 중립, 부정)를 이용하여 구축한 사전이다. 서술어는 형태소 분석기와 구분 분석기를 이용하여 자동으로 수집할 수 있지만, 평가 단어 사전을 구축하기 위한 단어의 의미와 평가 표현에 대한 판단은 사람에 의해 정의되어야 한다.

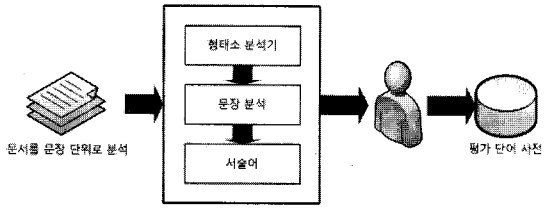

그림 4 평가 단어 사전 구축 과정
Fig. 4 Building process of reputation dictionary.

2.5. 적합성 (document relevance) 평가

단어의 연관도를 이용하여 문서를 분류하기 때문에 의미상 연결된 문서들끼리 분류를 할 수 있다. 문서 분류 과정을 거친 다음, 주제어에 의한 평가 표현으로 군집화하면, 주제어와 관련된 평가 표현을 가진 문서끼리 분류가 가능하다(그림 5).

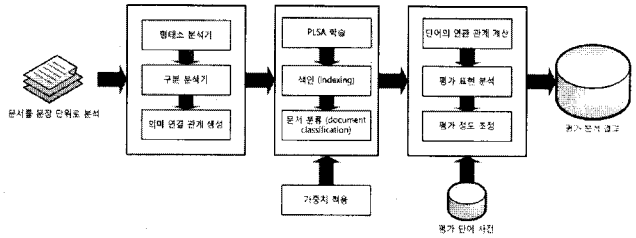

그림 5 PLSA 학습과정과 평가 분석 과정
Fig. 5 Process of PLSA learning and reputation analysis.

적합성 평가는 PLSA 학습 결과를 이용하여 단어 w_{query}나, 문서 d_{query}에 대한 카테고리 별 유사도를 평가하는 것이다. 적합성 평가의 결과에 의해 분석하려는 문서d_{query}나 단어 w_{query}가 어느 카테고리에 속하는 문서인지를 측정할 수 있다.

문서 d_{query}와 단어 w_{query}에 대한 적합성 평가는 다음과 같이 계산한다.

$$\text{문서의 적합성 평가: } P(d_i|d_{query}) = \sum_{z\in Z} P(d_i|z)P(z|d_{query})$$
$$P(z_n|d_{query}) = \frac{P(d_{query}|z_n)P(z_n)}{P(d_{query})} \quad (13)$$

$$\text{단어의 적합성 평가: } P(w_j|w_{query}) = \sum_{z\in Z} P(w_j|z)P(z|w_{query})$$
$$P(z_n|w_{query}) = \frac{P(w_{query}|z_n)P(z_n)}{P(w_{query})} \quad (14)$$

표 5는 카테고리 z_1과 문서 d_{query}의 적합성 여부를 계산한 결과이다. 적합성 평가를 통해 문서를 분류하거나, 해당 문서가 어떤 카테고리에 속하는지를 판단할 수 있다. 적절한 적합성 평가치는 실험적으로 결정해야 한다.

표 5 유사 클래스에 속한 문서 분류
Table 5 Document classification in similar class.

카테고리	단어 쌍	문서 #1	문서 #2	문서 #n
z_1	{LCD, 모니터}	0.93	0.93	…
	{모니터, 해상도}	0.81	0.87	
	{해상도, 높다}	0.63	0.8	
	{화질, 어둡다}	0.52	0.035	
	{화질, 만족한다}	0.62	0.756	
	{화질, 밝다}	0.38	0.34	
	…	…	…	
문서 평가치		0.75	0.43	…

3. 실험 및 결과

실험에 사용한 문서는 KRISTAL-IRMS의 KRTC-2003 한글 테스트 컬렉션[15]을 사용하였다. KRTC-2003은 120개의 카테고리 목록과 34,000개의 학습용 문서와 5,700개의 테스트 문서로 구성되어 있다.

KRTC-2003 한글 테스트 컬렉션을 그림 5와 같이 문서를 형태소 해석기와 구문 분석기를 이용하여 분석하고, 제안된 PLSA 알고리즘을 이용하여 색인, 문서 분류 과정을 거친 다음, 문서 내에서 주제어에 대한 평가 표현 단어와 구(句) 패턴을 추출하는 실험을 하였다. 잠재 의미 집합은 여러 실험을 통하여 k=150에서 가장 적절한 결과를 얻었으며, PLSA학습시 EM알고리즘의 반복 횟수는 70회에서 가장 좋은 성능을 얻었다. 문서의 카테고리별 유사도와 적합성 평가도는 실험을 통하여 0.7을 최저 평가치로 선택하였다. 먼저 학습용 문서를 대상으로 실험을 한 다음, 학습 결과를 통해 분류된 문서와 학습용 문서의 카테고리 목록을 비교하여, 카테고리별로 유사도를 비교하였다. 가중치를 적용한 PLSA 알고리즘이 기존 PLSA 알고리즘과 비교하여 약 10~15%의 향상된 결과를 보여 주었다(그림 6).

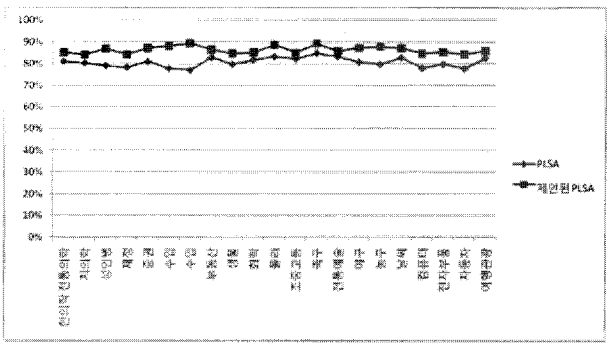

그림 6 카테고리 별 유사도 비교
Fig. 6 Comparison of similarity by categories.

그리고 학습 결과와 테스트용 문서를 각 카테고리 별로 문서 적합성 평가 실험을 하였다. 같은 카테고리에 속하는 테스트 문서는 약 70%이상의 문서가 평가 최저치를 만족하였다(그림 7).

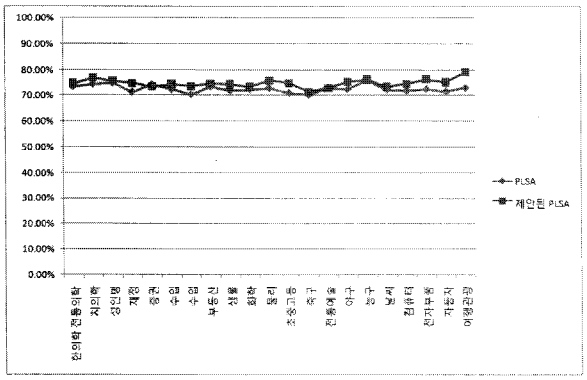

그림 7 문서 적합성 비교

Fig. 7 Comparison of document relevance.

실험에 사용된 문서에서 분석된 서술어는 3,152개이며, 평가 표현으로 사용되는 단어를 출현 빈도에 따라 상위 30개를 선택하여 평가 단어 사전를 구축하여, 평가 표현을 분석하였다.

표 6은 평가 표현 단어인 '저렴하다'를 기준으로 평가 표현을 분석한 결과이다.

표 6 평가 단어 분석

Table 6 Evaluation of reputation words.

주제어	평가단어	주제어
~공기를	저렴한	비용으로 공급하기위해 ~
승마기회를 ~ 비교적	저렴한	가격으로 회원모집에 ~
~ 승마장보다	저렴한	가입비를 ~
~ 값이	저렴한	것도 장점이다. ~
~ 더	저렴한	금리로 ~
~ 상상외로	저렴하다	
~ 수수료가	저렴하기	때문에 ~
~ 매매수수료도	저렴해	사이버거래가 ~
~ 금리가	저렴해	작년 ~
~ "작은화분 꽃시범판매장"을 설치	저렴한	가격으로
~ 가장	저렴한	수준으로
~ 영상회의시스템 보다	저렴하고	유지. 관리도 적은 ~

그림 8는 '저렴하다'와 관계된 주제어와 평가 표현의 관계를 추출하여 벡터 공간 모델로 표현한 결과이다. 두 단어 사이의 거리가 가까울수록 서로 연관성이 높은 단어로 볼 수 있다. 이 결과를 이용하여 '대출→{금리, 수수료}→{저렴하다, 높다, 비싸다}'와 같이 서로 의미가 연결된 연관 관계를 확인할 수 있다. '대출→{금리, 수수료}'는 서로 연관 단어로 볼 수 있다. 그리고,'대출→{금리, 수수료}'에 대해서는 '{비싸다, 높다}'보다 '{저렴하다}'라는 표현이 더 많이 사용된 것을 알 수 있다.

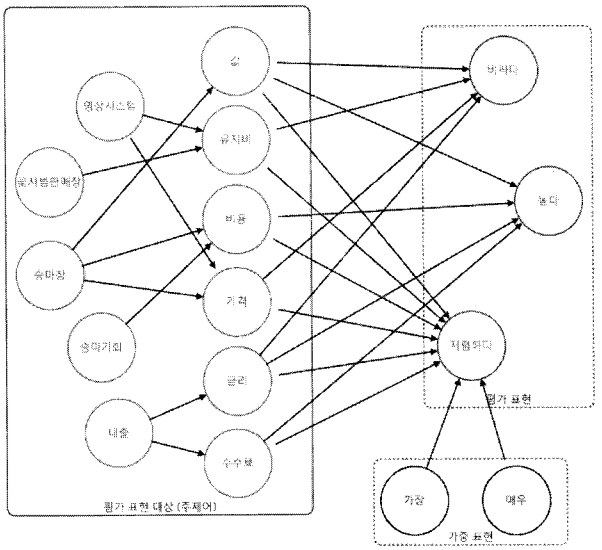

그림 8 주제어와 평가 표현 관계의 벡터 공간 모델

Fig. 8 Vector space model of the relationship between topic and reputation words.

4. 결 론

본 논문에서는 기존 키워드 검색이 가지고 있는 한계를 극복하기 위하여, PLSA 모델을 이용하여 단어의 의미 관계와 문서의 내용을 분석하였다. 실험을 통하여 서로 의미가 연결된 단어를 자동으로 분석하고, 연관 단어와 단어의 의미를 벡터 공간 모델로 표현하여, 기존 키워드 검색이 가지고 있는 단어의 의미 분석에 대한 한계를 극복할 수 있음을 확인하였다.

제안된 알고리즘은 다음과 같은 장점을 갖는다. 첫째, PLSA를 이용한 비감독 학습이기 때문에, 사전 정보와 문서간의 학습 정보가 필요 없다. 둘째, 구문 분석을 이용하여 주제어와 의미 표현의 관계를 백터 공간 모델로 표현하여, 주제어의 연관 단어와 이와 관련된 의미 표현 단어의 관계를 추출 할 수 있다. 셋째, 문서 분류와 적합성 평가를 이용하여, 의미 표현을 분석하기에 적합한 문서를 필터링 할 수 있다. 마지막으로 단일 문서뿐만 아니라 다중 문서에서도 특정 사건, 사물에 대한 의견이나 관심사를 분석할 수 있으며, 기존 키워드 검색의 한계를 극복하는데 도움이 될 것으로 기대한다.

평가 표현은 수치적으로 측정이 가능하지만, 평가 표현의 정도와 주제어에 대한 긍정, 부정 평가 표현을 정확하게 측정하기 위해서는 다양한 평가 표현에 대한 평가 단어 사전의 구축이 필요하다고 판단된다. 이 결과를 바탕으로 연관 단어 사전이나 평가 단어 사전을 구축하면 도움이 될 것이다. 앞으로 띄어쓰기와 같은 문법적 오류와 인용문과 같은 언어적 표현의 다양성을 극복하기 위한 연구가 진행 되어야 할 것이다.

참 고 문 헌

[1] Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2(1-2), pp. 1 - 135, 2008.

[2] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning, 42(1-2), pp. 177-196, 2001.

[3] Thomas Landauer, P. W. Foltz, and D. Laham, Introduction to Latent Semantic Analysis. Discourse Processes 25: 259 - 284, 1998.

[4] Daniel D. Lee and H. Sebastian Seung, Learning the parts of objects by non-negative matrix factorization, Nature, vol 401, pp. 788-791, 1999.

[5] 홍영국, 이종혁, 이근배, 의존문법에 기반을 둔 한국어 구문 분석기, 한국정보과학회 1993년 봄 학술논문발표집 제20권 제8호, pp. 33-46, 1994.

[6] P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1-21, 1977.

[7] R. J. Kozick and B. M. Sadler, Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures, IEEE Trans. on Signal Processing, vol. 48, No. 12, pp. 3520-3535, 2000.

[8] H. Chen, R. Perry, and K. Buckley, Direct and EM-based map sequence estimation with unknown time-varying channels, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2129-2132, 2001.

[9] R. A. Boyles, On the convergence of the EM algorithm, J. Roy. Sta. B., vol. 45, no. 1, pp. 47-50, 1983.

[10] C. Wu, On the convergence properties of the EM algorithm, Ann. Statist., vol. 11. 1, pp. 95-103, 1983.

[11] 김성수, 강지혜, 새로운 고속 EM 알고리즘, 한국정보과학회, 정보과학회논문지 : 시스템 및 이론 제31권 제9·10호, pp. 575-587, 2004.

[12] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, vol. 24, no. 5, pages 513-523, 1988.

[13] Shimodaira, H., Improving Predictive Inference under Covariate Shift by Weighting the Log-likelihood Function. Journal of Statistical Planning and Inference, Vol. 90, 227-244, 2000.

[14] 이경찬, 강승식, 범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템, 한국정보과학회, 한국정보과학회 2002년도 봄 학술발표논문집 제29권 제1호 (B), pp. 475 ~ 477, 2002

[15] 한국과학기술정보연구원, http://www.kristalinfo.com/K-Lab/Text-Cat/KRTC.2003.tar.gz

저 자 소 개

조 시 원 (曺 時 元)

1994년 동국대학교 전기공학과 (공학사). 1998년 동국대학교 대학원 전기공학과 (공학석사). 2003년~현재 동국대학교 대학원 전기공학과 박사과정

Tel : 02-2260-3350

E-mail : stsolaris@gmail.com

이 동 욱 (李 東 旭)

1983년 서울대학교 전기공학과 (공학사). 1985년 서울대학교 대학원 전기공학과 (공학석사). 1992년 Georgia Tech.대 (공학박사). 1993년~현재 동국대학교 전기공학과 교수

Tel : 02-2260-3350

E-mail : dlee@dongguk.edu