마코프 논리 기반의 시맨틱 문서 검색

(Semantic Document-Retrieval Based on Markov Logic)

황 규 백 † 봉 성 용 ††
(Kyu-Baek Hwang) (Seong-Yong Bong)
구 현 서 ††† 백 은 옥 ††††
(Hyeonseo Ku) (Eunok Paek)

요 약 본 논문은 질의 문서와 의미가 유사한 문서를 검색하는 문제를 다룬다. 이 문제에 대한 기본적인 접근법은 각 문서를 bag-of-words 형태로 표현한 후, 코사인 유사도 등의 거리 기준에 기반하여 유사 문서를 판별하는 것이다. 그러나, 이처럼 문서에 출현하는 단어에만 의존하는 검색 방법은 의미적 유사성을 제대로 반영하기 어렵다는 단점을 가진다. 본 논문에서는 이러한 문제를 극복하기 위해 데이터 기반의 감독 학습(supervised learning) 기법과 관련 온톨로지 정보를 마코프 논리(Markov logic)에 기반하여 결합한다. 구체적으로, 단어들 사이에 존재하는 관계를 표현한 온톨로지와 유사도가 태깅된 문서 데이터에서 마코프 논리 망(Markov logic network)을 학습하며, 학습된 마코프 논리 망과 문서 데이터 및 새로 주어진 질의 문서에 대한 추론을 통해 질의 문서와 의미적으로 유사한 문서를 검색하는 기법을 제안한다. 제안하는 접근법은 서울시의 민원서비스 홈페이지에서 수집된 실제 민원 데이터에 적용되었으며, 적용 결과, 단순한 문서 간 거리에 기반한 유사 문서 검색 기법에 비해 월등히 높은 정확도를 보였다.

키워드 : 정보검색, 시맨틱 문서 검색, 감독학습, 온톨로지, 마코프 논리

Abstract A simple approach to semantic document-retrieval is to measure document similarity based on the bag-of-words representation, e.g., cosine similarity between two document vectors. However, such a syntactic method hardly considers the semantic similarity between documents, often producing semantically-unsound search results. We circumvent such a problem by combining supervised machine learning techniques with ontology information based on Markov logic. Specifically, Markov logic networks are learned from similarity-tagged documents with an ontology representing the diverse relationship among words. The learned Markov logic networks, the ontology, and the training documents are applied to the semantic document-retrieval task by inferring similarities between a query document and the training documents. Through experimental evaluation on real world question-answering data, the proposed method has been shown to outperform the simple cosine similarity-based approach in terms of retrieval accuracy.

Key words : information retrieval, semantic document-retrieval, supervised learning, ontology, Markov logic

1. 서 론

유사 문서 검색은 사용자가 질의하는 문서와 의미적으로 유사한 문서를 문서 데이터베이스에서 검색하는 것을 의미한다[1,2]. 이러한 유사 문서 검색은 논문 및 특허의 표절 여부(혹은 유사도)를 평가하는 시스템이나 사용자가 질의하는 민원과 유사한 기존 민원을 검색하여 관련 답변을 알려주는 서비스 등에서 유용하게 사용될 수 있다. 이는 질의어가 단어나 구 혹은 문장 수준이 아니라, 검색되어야 할 문서와 비슷한 크기의 문서라는 점에서 기존의 질의어에 기반한 웹 검색 등과 구별된다.

기본적인 접근법은 문서를 bag-of-words 형태로 표현하고 그 유사도를 계산하는 것이다. 다만, 질의가 문서 자체이기 때문에 몇 가지 문제점들을 고려해야 한다. 첫째, 질의 문서에는 질의와는 상관이 없는 단어들이 상당수 존재한다. 둘째, 문서 자체가 두 개 이상의 주제를 가질 수 있다. 셋째 문제는 다른 경우의 검색에도 공통적으로 적용될 수 있는 사항인데, 단어만 가지고 문서의 유사도를 평가하는 경우, 각 단어가 내포하는 배경 지식이 고려되지 않기 때문에, 의미적으로 볼 때 유사하다고 판단할 수 없는 결과가 나올 가능성이 있다. 위의 문제들 중 두 번째 문제를 해결하기 위해 [3,4]는 문서를 다수의 부 주제(subtopic)로 나누었다.

· 본 논문은 서울시정개발연구원의 서울시 기반기술구축사업(GS070167)의 지원으로 연구되었음. 황규백, 봉성용은 숭실대학교 교내연구비의 지원을 받았음.
· 이 논문은 제36회 추계학술발표회에서 '마코프 논리 기반의 시맨틱 문서 검색'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : 숭실대학교 컴퓨터학부 교수
kbhwang@ssu.ac.kr
†† 학생회원 : 숭실대학교 컴퓨터학과
sybong@ml.ssu.ac.kr
††† 비 회 원 : 서울시립대학교 기계정보공학과 연구원
yorq@paran.com
†††† 종신회원 : 서울시립대학교 기계정보공학과 교수
paek@uos.ac.kr

논문접수 : 2009년 12월 23일
심사완료 : 2010년 3월 11일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제6호(2010.6)

본 논문에서는 첫째, 셋째 문제를 완화하기 위해, 감독 학습(supervised learning) 기법과 온톨로지 정보를 결합한다. 구체적으로, 유사도가 태깅된 문서 데이터에 감독 학습 기법을 적용하며, 이때 별도로 구성된 온톨로지 정보를 함께 활용한다. 이를 위해 마코프 논리(Markov logic)를 적용한다[5]. 제시하는 기법은 서울시에서 수집한 약 4,000 건의 민원 문서에 적용되었으며, 기본적인 코사인 유사도 기반의 유사 문서 검색에 비해 그 정확도가 월등히 향상됨을 실험을 통해 보인다. 논문의 구성은 다음과 같다. 2절에서는 마코프 논리에 대해 소개하고, 3절에서는 본 논문에서 제안하는 유사 문서 검색 방법을 기술한다. 4절에서는 실제 데이터에 제안한 방법을 적용한 실험 결과를 보이며, 마지막으로 5절에서 결론 및 향후 연구 방향을 제시한다.

2. 마코프 논리(Markov Logic)

2.1 표현 및 모델 정의

본 절에서는 마코프 망(Markov network)에 대해 간략히 기술한 후, 마코프 논리를 설명한다. 마코프 망은 변수들의 결합확률분포를 나타내는 모델로 무방향 그래프(undirected graph) G와 포텐셜 함수들 Φ_k로 구성된다. (k는 망 구조 그래프에서의 클릭(clique)을 가리키는 지수(index)이다.) 이 때, 변수집합 X에 대한 마코프 망이 나타내는 결합확률분포는 다음과 같이 표현된다.

$$P(X=x)=\frac{1}{Z}\prod_k \Phi_k(x_{\{k\}}) \tag{1}$$

여기서 $x_{\{k\}}$는 k번째 클릭의 상태(configuration)이며, 분할 함수(partition function) Z는 아래와 같이 정의된다.

$$Z=\sum_{x\in X}\prod_k \Phi_k(x_{\{k\}}) \tag{2}$$

마코프 망을 로그 선형 모델(log-linear model)을 이용해 나타내면, 포텐셜 함수가 상태의 가중치의 합을 나타내는 함수로 바뀌며, 식 (1)은 다음과 같이 변경된다.

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_j w_j f_j(x)\right) \tag{3}$$

위 식에서 j는 클릭을 나타내는 지수이고, $f_j(x)$는 해당 클릭의 상태에 대해 정의된 자질(feature)의 함수이며, w_j는 해당 클릭에 대한 가중치이다.

한편, 1차 논리의 지식 베이스는 참인 공식(formula)들의 집합이며, 가능한 세계(possible world)에 대해 제약을 가하는 것으로, 어떤 세계가 공식을 하나라도 위반한다면, 그 세계의 확률은 0이 된다. (즉, 거짓이 된다.) 마코프 논리는 이 규칙을 유연하게 적용해 0~1 사이의 값으로 각 세계의 확률를 표현한다. 이를 위해 각 공식은 가중치를 가지며, 이는 그 공식이 얼마나 강한 제약인지 나타낸다. 전술한 마코프 망은 이를 위해 활용되며, 논리와 마코프 망이 결합된 마코프 논리 망(Markov logic network, MLN)은 아래와 같이 정의된다[5].

정의. 마코프 논리 망(Markov logic network) L은 (F_i, w_i)의 집합이다. 여기서 F_i는 1차 논리의 공식이며, w_i는 실수값으로 주어지는 가중치이다. L은 유한한 상수 집합 $C = \{c_1, c_2, \ldots, c_{|C|}\}$와 함께 마코프 망 $M_{L,C}$를 다음과 같이 정의한다.

1. $M_{L,C}$는 L의 각 술어의 모든 가능한 기저 예(ground instance 혹은 grounding)에 대응하는 노드를 가지며, 그 값은 기저 술어가 참이면 1, 거짓이면 0이다.
2. $M_{L,C}$는 F_i의 모든 가능한 기저 예에 대해 자질을 가지며, 그 값은 기저 공식(ground formula)이 참일 때 1, 아니면 0이다. w_i는 이 자질의 가중치이다.

정리하자면, 각 술어의 인자에 상수를 적용한 것이 각 노드를 구성하며, 하나의 공식을 함께 구성하는 노드들끼리 마코프 망에서 연결된다. 이러한 마코프 망의 결합분포는 식 (3)에 기반하여 다음과 같이 표현된다.

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_{i=1}^{F} w_i n_i(x)\right) \tag{4}$$

여기서, F는 공식의 개수이며, $n_i(x)$는 주어진 세계 x에서 F_i가 참인 경우의 수이다.

2.2. 추론

마코프 논리에서의 추론은 주어진 근거(evidence)에 기반해 참인 공식(마코프 망에서 클릭에 해당)들의 가중치의 합을 가장 크게 만드는 진리 값 할당을 찾는 것이다. 주어진 근거가 없을 때 가장 그럴듯한 세계(most probable explanation, MPE)를 찾는 작업은 weighted satisfiability 문제와 동일하므로, SAT solver 알고리즘이 사용될 수 있다[5]. 근거가 주어졌을 때 특정 공식의 확률 계산은 marginalization을 요구하기 때문에, 주로 근사 기법인 마코프 체인 몬테 카를로(Markov chain Monte Carlo, MCMC) 방법이 적용된다[5].

2.3 학습

마코프 논리 망(MLN) 학습은 구조 및 가중치 학습으로 나뉘며, 본 절에서는 가중치 학습에 대해 기술한다. 이는 1차 논리의 기저 예들의 진리값이 주어진 경우, 이를 가장 잘 표현하는 가중치를 계산하는 것으로, 아래 식과 같이 gradient ascent에 기반하여 우도(likelihood)를 최대화하는 가중치를 구하는 것과 동일하다.

$$\frac{\partial}{\partial w_i}\log P_w(X=x)=n_i(x)-\sum_{x'}P_w(X=x')n_i(x') \tag{5}$$

위에서 우변의 합은 모든 가능한 진리값의 경우 x'에 대한 것이며, $P_w(X=x')$은 현재의 가중치 벡터 $w = (w_1, \ldots, w_i, \ldots)$를 이용해 $P(X=x')$을 계산한 것이다. 이 식의

계산은 셈 문제(counting problem)로 그 복잡도가 #P-complete에 해당하며, 보통은 의사 우도(pseudolikelihood)나 조건부 우도(conditional likelihood)를 최적화한다. 유사 문서 검색 문제에는 문서의 유사도라는 목표 값(target value)이 있기 때문에, 조건부 우도를 최적화하는 방법을 적용하며, 이는 다음과 같다.

조건부 우도 최적화에서는 목표 값에 해당하는 기저 원자(ground atom)를 질의 원자(query atom)라 하고 나머지를 근거 원자(evidence atom)라 한다. 근거 원자를 X, 질의 원자를 Y라 하면, X가 주어질 때 Y의 조건부 우도에 대한 gradient는 다음과 같다[5].

$$\frac{\partial}{\partial w_i}\log P_w(y|x) = n_i(x,y) - \sum_{y'} P_w(y'|x) n_i(x,y') = n_i(x,y) - E_w[n_i(x,y)] \quad (6)$$

위에서 기대값 $E_w[n_i(x,y)]$의 계산은 많은 계산량을 요구하므로, MAP(maximum a posteriori) 상태 $y^*_w(x)$에서 $n_i(x,y^*_w)$를 세어 근사값으로 쓰는 방법을 취한다.

3. 마코프 논리 기반 유사 문서 검색 기법

2절에서 설명한 마코프 논리에 기반한 유사 문서 검색 과정은 그림 1과 같다. 우선, 각 문서를 전처리하여 bag-of-words 표현으로 변형한 뒤, 군집화한다. 이는, 마코프 논리 망에서의 추론 및 학습이 요구하는 방대한 계산량에서 초래되는 문제를 완화하는 효과와 함께 감독 학습에 필요한 정답 집합 구축을 가능케 하는 효과[1]를 가진다.

군집 내부의 문서 쌍에 대해 유사도 태깅이 이루어지면 마코프 논리 학습을 위한 데이터가 구축된다. 이 학습 데이터와 이미 구축된 온톨로지에 기반하여 각 군집별로 마코프 논리 망의 가중치가 학습된다. 이 마코프 논리 망은 유사 문서 검색을 위한 중요도가 부착된 1차 논리 규칙 베이스로 볼 수 있다.

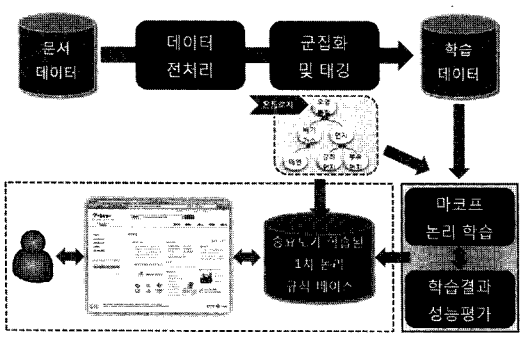

그림 1 마코프 논리 기반 유사 문서 검색 개요

이후, 질의 문서가 입력되면 해당 군집을 결정하고 그 군집에 해당하는 마코프 논리 망에 질의 문서 및 온톨로지를 결합하여, 추론을 통해 군집 내에서 유사 문서를 검색한다. 다음 절에서는 이를 보다 상세히 기술한다.

3.1 전처리, 군집화 및 학습 데이터 구축

각 문서마다 형태소 분석을 통하여 적절한 빈도의 명사를 추출한 뒤, 각각의 TF/IDF 값을 계산한다. TF(term frequency)는 특정 문서에서의 특정 단어의 빈도수에 해당하며, IDF(inverted document frequency)는 특정 단어가 출현한 문서의 전체 문서 집합에서의 빈도수에 로그를 취한 값이 널리 쓰인다. 이렇게 전처리 과정을 거친 데이터에 대해 k-평균(k-means) 군집화 기법을 적용한다. 이후 각 군집 내부의 문서 쌍에 대해 유사도 여부를 태깅하여 학습 데이터를 구축한다.

3.2 각 군집 별 마코프 논리 망 학습

문서 및 온톨로지를 기술하기 위해 본 논문에서 정의한 술어의 예는 다음 표와 같다.

```
HasWord(document,word)
SubClassOf(word,word)
SimPet(document,document)
```

HasWord는 특정 문서가 특정 단어를 가지는 사실을 나타내는 술어이며, SubClassOf는 본 논문에서 적용한 온톨로지 중 단어들 사이의 상하위어 관계를 표시하는 술어이다.[2] SimPet은 두 문서가 유사한지의 여부를 나타내는 술어로 마코프 논리 적용 시, 질의 술어에 해당한다.

마코프 논리 망 구성을 위해 사용된 공식 중 온톨로지와 상관이 없는 것은 아래와 같다.

```
HasWord(d1,+w) ^ HasWord(d2,+w) => SimPet(d1,d2)
HasWord(d1,w1) ^ HasWord(d2,w2) ^ (d1 != d2) ^
    (w1 != w2) => !SimPet(d1,d2)
HasWord(d1,+w) ^ !HasWord(d2,+w) ^ (d1 != d2) =>
    !SimPet(d1,d2)
```

위의 공식들은 문서 사이의 관계에 대한 일반적인 사실들이다. 예를 들어, 마지막 공식은, 특정 단어를 문서 d1은 가지고 있고, 문서 d2는 가지고 있지 않다면, 둘은 유사하지 않다는 사실을 기술하고 있다. 일반적인 1차 논리에서는 위의 공식들이 참 또는 거짓인 둘 중 하나의 진리값을 가지지만, 마코프 논리에서는 가중치를 가질 수 있다. 위의 상수 표현에서 +기호는 각 상수가 적용되어 만들어질 수 있는 규칙들에 대해서 가중치를 따로 학습해야 함을 나타낸다. 예를 들어, 마지막 공식은 각 단어들에 대해서 따로 가중치가 계산된다.

아래 표는 온톨로지 관계를 적용하기 위한 규칙들의 일부를 나타낸다.

1) 유사 문서 정답 집합 구축을 위해서는 문서 개수의 제곱에 해당하는 경우의 수를 고려해야 한다.

2) 단어들 사이의 관계는 상하위어 외에 다른 것들도 존재하며, 본 논문에서 사용한 온톨로지는 총 601개의 단어에 대해 구축되었다.

```
HasWord(d1,HaSuDo) ^ HasWord(d2,SiSeol) ^
SubClassOf(HaSuDo,SiSeol) => SimPet(d1,d2)
HasWord(d1,GiChe) ^ HasWord(d2,MulJilSangTai) ^
SubClassOf(GiChe,MulJilSangTai) => SimPet(d1,d2)
HasWord(d1,NongOyag) ^ HasWord(d2,OoOyeomMulJil)
^ SubClassOf(NongOyag,OoOyeomMulJil) =>
SimPet(d1,d2)
HasWord(d1,SeogOyuGiGwan) ^
HasWord(d2,NaiOyeonGiGwan) ^
SubClassOf(SeogOyuGiGwan,NaiOyeonGiGwan) =>
SimPet(d1,d2)
...
```

학습이 끝나면, 위의 공식들은 주어진 학습 데이터를 가장 잘 표현할 수 있는 가중치를 가지게 된다.

3.3 마코프 논리에 기반한 추론 및 민원 검색

질의 문서가 속하는 군집을 찾기 위해, 각 군집의 중심점과의 코사인 유사도를 계산한다. 중심점은 소속 문서들의 TF/IDF 벡터 값의 평균이다. 선정된 군집의 학습 데이터에 질의 문서의 HasWord 정보를 추가하여 마코프 논리를 적용하기 위한 지식 베이스를 만든다. 이 지식 베이스는 추론 시 근거에 해당한다. 이후, 주어진 근거를 바탕으로 학습 데이터 내부의 각 문서와 질의 문서 사이의 SimPet 술어의 확률을 마코프 논리 망에서의 추론을 통해 계산한다. SimPet 확률이 높은 문서를 질의 문서와 유사한 문서로 하여 검색 결과를 도출한다.

추가적으로, 추론 시에는 학습과 상관없이 항상 참인 아래의 공식[hard constraint]을 추가할 수 있다.

```
SimPet(d,d)
SimPet(d1,d2) => SimPet(d2,d1)
```

4. 실험 및 결과

4.1 실험 설정

제안한 방법을 2007, 2008년 서울시의 12개 부서에서 수집한 4041건의 민원 데이터에 대해 적용하였다. 민원은 제목, 민원 내용, 답변으로 구성되어 있으며, 본 연구에서는 민원 내용을 유사 문서 검색의 대상으로 적용하였다. 형태소 분석기는 ㈜솔트룩스의 형태소 분석기를 활용하였으며, 빈도수가 20 이하인 단어들은 고려 대상에서 제외하였다. k-평균 군집화의 k 값으로는 다양한 값을 적용한 뒤, 태깅과 학습 및 추론에 적절한 크기의 군집이 배출된 110을 활용하였다. 110개의 군집 중 포함하는 문서의 개수가 4~83인 67개의 군집의 문서를 태깅하였다.[3] 67개의 군집에 포함된 문서의 개수는 약 2000개이다. 온톨로지 및 문서 기술을 위해 적용된 술어는 모두 68개이며, 마코프 논리 망 학습을 위해 사용된 공식은 온톨로지와 관계없는 것이 5개, 온톨로지를 고려한 규칙이 294개였다.

마코프 논리의 적용은 공개 소프트웨어인 Alchemy (http://alchemy.cs.washington.edu)를 활용하였다. 학습은 조건부 우도를 최대화하는 구분 학습(discriminative learning) 기법을 적용하였으며, 제안하는 방법의 성능을 상대적으로 평가하기 위해 각 군집에 대한 코사인 유사도 기반의 검색 방법[6]과 나이브 베이즈 분류기[7] 방법[4]을 활용하였다.

정확도 평가는 각 군집의 문서 중 70%를 학습에 이용하고, 나머지 30%의 문서를 질의로 한 경우의 F1-measure에 기반하였다.

4.2 실험 결과

아래의 표 1은 제안 기법(MLN), 코사인 유사도 기반(Cosine), 나이브 베이즈 분류기 기반(NBC)의 67개 군집에 대한 평균 정확도를 비교하고 있다.

표 1 제안 기법(MLN), 코사인 유사도 기반 기법(Cosine) 및 나이브 베이즈 분류기 기반 기법(NBC)의 검색 정확도 비교 (F1-measure)

	MLN	*Cosine*	*NBC*
평균	0.6935528	0.6180089	0.6002856
표준편차	0.344735	0.333642	0.3265152

결과를 보면, 제안하는 기법의 정확도가 기존의 방법 및 나이브 베이즈 분류기 기반의 기법에 비해 평균적으로 뛰어남을 알 수 있다. 제안하는 기법의 성능이 많이 우수하지 못한 경우는, 주로 군집 내부의 거의 모든 문서가 유사하다고 태깅이 되어 있거나, 혹은 반대로 거의 모든 문서쌍이 유사하지 않다고 태깅이 되어 있는 경우였다. 이러한 경우는 심각한 데이터 불균형 문제 때문에 감독 학습 기반의 방법이 정확한 결과를 내기 어려우며, 굳이 감독 학습을 이용할 필요 없이 단순한 문서 간 유사도 기반의 방법을 적용해도 무방한 경우라고 볼 수 있다.

다음은, 제안하는 방법의 어느 요소가 정확도 향상에 기여하는지 알아보기 위해, 군집 데이터가 너무 불균형에 치우치지 않은 경우만을 대상으로, 1) 온톨로지 정보의 영향 및 2) 학습 후 추가된 항상 참인 공식(3.3절 참조)의 영향을 평가하였다. 그림 2, 3, 4는 23개의 특정 군집(소속된 문서 개수가 21 이상이고, 문서쌍 중 유사한 것으로 태깅된 것의 비율이 0.05 이상 0.8 미만)에서 두 가지 요소의 조합에 따른 성능 추이를 보이고 있다. 구체적으로, Default는 온톨로지 정보와 항상 참인 공식을 제거한 경우, Ontology는 온톨로지 관련 공식만을

3) 문서의 개수가 너무 적은 경우는 태깅의 의미가 없어서 제외하였고, 너무 많은 경우는 고려해야 할 문서쌍이 너무 많아져서 제외하였다.

4) 이는 태깅된 문서를 감독 학습의 데이터로 활용하여 나이브 베이즈 분류기를 온톨로지 정보 없이 적용한 것이다.

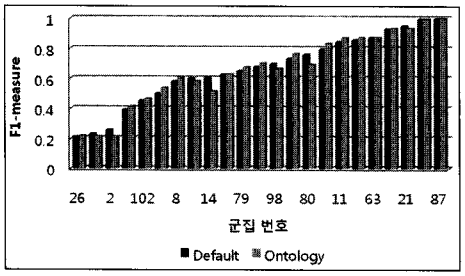

그림 2 온톨로지 정보의 영향

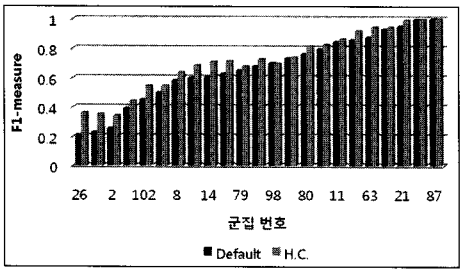

그림 3 항상 참인 공식(hard constraint)의 영향

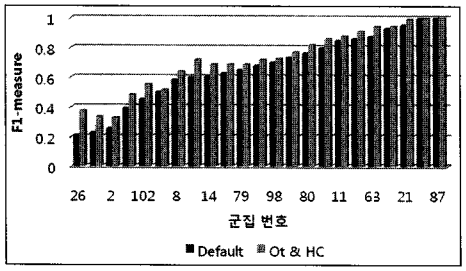

그림 4 온톨로지 및 항상 참인 공식의 영향

추가한 경우, H.C.는 항상 참인 공식만을 추가한 경우, Ot & HC는 두 정보를 모두 활용한 경우를 나타낸다.

그림 2, 3, 4를 보면, 마코프 논리에 기반한 방법에 있어서 온톨로지 정보와 항상 참인 공식이 어떤 영향을 미치는지 확인할 수 있다. 온톨로지 정보만을 적용했을 경우 F1-measure의 군집별 평균은 0.6623에서 0.6621로 미세하게 떨어졌지만, 23개 중 16개의 군집에서 성능 향상을 보였다. 항상 참인 규칙을 추가했을 경우 평균 정확도는 0.716으로 뚜렷한 증가를 보였으며, 긍정적 효과를 본 군집의 개수도 22개로 늘어났다. 온톨로지 정보와 항상 참인 공식을 동시에 적용했을 경우에는 F1-measure의 평균이 0.721로 크게 증가했으며, 모든 군집들에서 성능 향상을 확인할 수 있었다.

5. 결 론

본 논문에서는 감독 학습과 온톨로지를 결합하여 시맨틱 문서 검색의 성능을 획기적으로 높일 수 있는 기법을 제안하였다. 제안한 기법은 1차 논리의 표현력과 확률그래프모델의 유연성을 동시에 가지고 있는 마코프 논리에 기반하고 있다. 구체적으로, 온톨로지 및 학습 데이터를 1차 논리의 학습 데이터로 나타내고, 문제를 서술하는 규칙의 가중치를 그 학습 데이터에 기반하여 학습하였다. 이후, 새로운 문서가 주어졌을 때 학습된 가중치를 가지는 규칙 베이스에서의 적절한 추론을 통하여 그 문서와 유사한 문서를 검색할 수 있었다. 실제 데이터에 대한 실험에서, 제안된 기법은 기존의 방법에 비해 검색 정확도의 향상을 가져왔으며, 이는 특히 기존의 단어 기반 유사도로는 유사 여부를 판단하기 어려운 경우에 두드러졌다. 향후 연구 방향은 온톨로지 및 다양한 정보가 성능 향상을 가져오는 원인을 보다 상세히 분석하여 제시된 기법을 고도화하는 것과 태깅의 어려움을 극복할 수 있는 전이 학습(transfer learning) 기법 등을 적용하는 것을 들 수 있다.

참 고 문 헌

[1] Atlam, E., Fuketa, M., Morita, K., and Aoe, J., Documents similarity measurement using field association terms, *Information Processing and Management*, vol.39, no.6, pp.809-824, 2003.

[2] Saracoglu, R., Tuetuencue, K., and Allahverdi, N., A fuzzy clustering approach for finding similar documents using a novel similarity measure, *Expert Systems with Applications*, vol.33, no.3, pp. 600-605, 2007.

[3] Takaki, T., Fujii, A., and Ishikawa, T., Associative document retrieval by query subtopic analysis and its application to invalidity patent search, *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pp.399-405, 2004.

[4] Wan, X., Yang, J., and Xiao, J., Towards a unified approach to document similarity search using manifold-ranking of blocks, *Information Processing and Management*, vol.44, no.3, pp.1032-1048, 2008.

[5] Domingos, P. and Lowd, D., *Markov Logic: An Interface Layer for Artificial Intelligence*, Morgan & Claypool, 2009.

[6] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, ACM Press and Addison Wesley, 1999.

[7] Domingos, P. and Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, vol.29, pp.103-130, 1997.