WordNet과 BPN을 이용한 웹 문서 적합성 판단

김원우[0], 변영태
홍익대학교 컴퓨터공학과
wwkim@cs.hongik.ac.kr byun@cs.hongik.ac.kr

Deciding The Relevance of Web Documents Using WordNet and BPN

Won-Woo Kim[0], Young-Tae Byun
Dept. of Computer Science, Hongik University

요 약

본 논문은 웹 문서가 특정 주제와 관련된 정보를 담고 있는지를 특정 주제의 단어와 다른 주제의 단어들 사이의 관계를 이용해 평가할 수 있는 방법을 제시하고자 한다. 특정 주제와 관련된 웹 문서에 단어$_A$와 단어$_B$가 그렇지 않은 웹 문서보다 나온 수가 더 많다면, 단어$_A$와 단어$_B$의 연결 관계는 특정 주제에 대해 Positive하다고 볼 수 있다. 반대의 경우에는 Negative하다고 볼 수 있다. 이러한 단어와 단어의 연결 관계를 수치화하여 특정 주제와 관련된 웹 문서의 평가에 사용할 수 있도록 WordNet과 BPN을 이용해 보고자 한다.

1. 서론

본 연구실에서 특정 주제 '동물'에 대해 사용자에게 정보와 웹 문서를 제공해 주는 Web Information Agent(HIIA)의 개발 중에 있다[1]. Web Information Agent는 동물 정보를 담고 있는 웹 문서 서비스를 위해, 미리 관련문서들(동물과 관련된 정보를 담고 있는 웹 문서)을 인터넷에서 수집해 저장해 놓는다. Web Information Agent에서 관련문서의 수집을 담당하는 Web Manager는 웹 문서를 여러 단계의 조건들을 거쳐 평가한다. 문서를 평가하는 조건들을 연구하면서[2], 동물 관련 단어가 어떤 단어와 함께 사용되었느냐가 관련문서 평가의 주요 조건이 될 수 있다는 점을 발견하였다.

예를 들어, shark와 food가 함께 나온 경우와 chicken이 food와 함께 나온 경우를 보자. shark-food에서 food는 상어가 먹는 '먹이'로, chiken-food에서 food는 닭의 '먹이'라는 의미 외에도 닭을 재료로 한 '음식'의 의미로도 해석될 수 있다. 같은 food라도 어떤 동물 관련 단어과 함께 사용되었는지에 따라 '먹이'라는 동물과 관련된 의미가 되기도 하지만, '음식'이라는 상관없는 의미가 되기도 한다.

본 논문에서는 몇 개의 선택된 주제에 포함된 단어들에 대하여 단어와 단어 간의 관계를 수치화하고, 이들을 가지고 주제별로 구해진 값들을 BPN에 Input vector로 입력하여, 동물 정보를 가진 웹 문서들을 평가할 수 있도록 학습시키는 방법에 대해 소개하고자 한다.

논문의 2절에서는 주제별 단어 그룹 구성에 사용한 WordNet에 대해서 살펴본다. 3절은 BPN, 4절은 학습 및 테스트, 5절은 결과에 대해서 기술하고, 마지막으로 6절에서 결론 및 향후과제를 맺고자 한다.

2. WordNet 소개와 이용

2.1 WordNet에 대하여

Ontology의 일종으로 간주되고 있는 WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤 대학 인지과학 연구실이 구축해온 언어어휘 데이터베이스이다[3][4]. WordNet의 주된 특징은 단어형이 아닌 단어의 의미를 구성 요소로 하였다는 점이다. WordNet은 현재 자연언어처리 및 정보검색의 여러 분야에서 널리 이용되고 있으며, 다국어판으로 번역되어 사용된다.

2.1 선택된 주제들

단어와 단어의 관계에서 모든 가능한 단어들을 연결한 경우의 수는 무한하기 때문에, 실험에서는 동물과 몇 개의 주제들을 선택해 여기에 속한 단어들끼리만 연결된 경우로 그 수를 한정하고자 한다. 동물 외의 주제들은 검색엔진들(altavista, excite, lycos)이 나누어 놓은 카테고리로부터 44개를 선별했고, art, book, business, car 등이 포함되어 있다(<표 1>참조). 모든 웹 문서는 이 주제들 중에 하나로 완벽하게 평가되거나, 포괄적으로 어떤 주제에 속하는 관계로 평가될 수 있을 것이다.

2.2 WordNet을 이용한 주제별 단어 그룹의 생성

주제는 의미적으로 그 주제와 관련된 단어들을 모아놓은 형태로 표현하고자 한다. 주제와 관련된 단어들은 WordNet (영문 버전)을 이용해 수집하고자 한다.

```
animal, animate being, beast, brute, creature, fauna -- (a living organism characterized by voluntary movement)
  => vermin -- (any of various small animals or insects that are pests; e.g. cockroaches or rats)
  => varmint, varment -- (any usually predatory wild animal considered undesirable; e.g. coyote)
  => scavenger -- (any animal that feeds on refuse and other decaying organic matter)
  .. 이하 생략
```

<예제 1> 'animal'로 가져온 WordNet 자료 (Noun Hyponyms:full)

<예제 1>은 동물 주제를 대표하는 'animal'을 WordNet에 Search Word로 입력하여 출력된 결과이다. 첫 줄은 'animal'과 그의 동의어들이 나오고, '=>'은 'animal'과 관련된 단어들을 계층적으로 관계를 표현해 주고 있다. 위와 같은 방법으로 모든 주제별로 자료를 수집하였고, 계층 관계를 무시한 단어 그룹을 만들었다. 주제별 단어의 수는 <표 1>과 같다.

본 연구는 뇌과학 연구 사업의 지원으로 진행 되었음.

2718	413	189
564	54	11
10	19	50
7	19	85
9	27	11
7	31	292
18	17	42
19	9	117
117	15	185
91	9	12
67	29	247
21	27	210
34	15	21
482	101	119
44	21	322

<표 1> 주제별 영문 명사 단어의 수
(CS: Computer Science)

3. BPN (Back-Propagation Network)

3.1 $word_a$:$subject_s$.$word_b$

단어와 단어의 관계를 이용해 웹 문서가 동물 관련 정보를 담고 있는 문서인지를 학습시키기 위해서, 먼저 단어와 단어사이의 관계를 수치화하는 과정이 필요하다. 수치화된 단어와 단어의 관계를 이용해 BPN의 Input layer에 들어가는 값들이 계산된다. $word_a$:$subject_s$.$word_b$는 동물 주제에 속한 $word_a$와 $subject_s$에 속한 $word_b$의 연결사이의 수치화된 값을 의미한다. 이 값은 다음과 같이 구한다.

$$word_a : subject_s.word_b = \alpha \times \frac{RE(word_a, subject_s.word_b) - IR(word_a, subject_s.word_b)}{RE(word_a, subject_s.word_b) + IR(word_a, subject_s.word_b)}$$

$RE(word_a,subject_s.word_b)$: $word_a$와 $subject_s$의 $word_b$가 Training Set의 관련문서들 중에 같이 나온 문서 수
$IR(word_a,subject_s.word_b)$: $word_a$와 $subject_s$의 $word_b$가 Training Set의 비관련문서들 중에 같이 나온 문서 수

이렇게 하면 Training Set의 관련문서에 많이 나온 $word_a$:$subject_s$.$word_b$의 값은 $+\alpha$에 가까워 지고, 반대로 비관련문서에 많이 나온 경우에는 $-\alpha$에 가까워 진다. 예를 들어, α가 1인 경우, RE(shark,food.food)가 11이고, IR(shark, food.food)가 4라면,

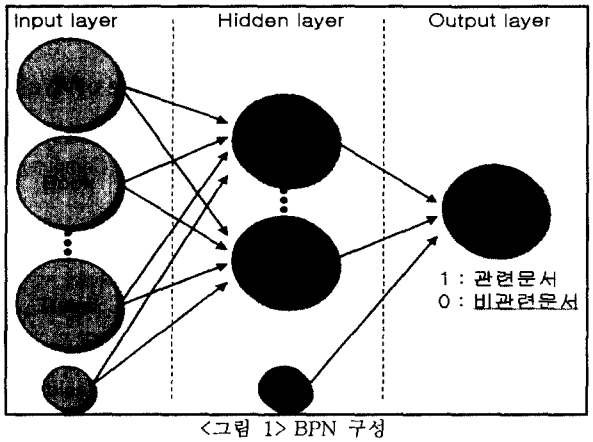
$$\text{shark:food.food} = 1 \times \frac{11-4}{11+4} = +0.467$$

동물 관련 단어인 shark와 주제 food의 food는 서로 +0.467의 관계(positive relationship)를 가진다고 보며, shark와 food가 같이 나온 웹 문서의 경우에 좋은 동물 관련문서일 가능성이 높다.

$word_a$:$subject_s$.$word_b$ 값이 0이 나온 경우는 계산 비용의 절감을 위해서 제거하였고, $RE(word_a,subject_s.word_b)$ 또는 $IR(word_a,subject_s.word_b)$가 1인 경우에는 정확도가 떨어진다고 보고 제거하였다. 이렇게 하여 총 27,421개의 $word_a$:$subject_s$.$word_b$ 값들을 얻을 수 있었다.(α 값은 1로 하였다)

3.2 BPN의 구현

BPN은 <그림 1>과 같이 구현하였고, 학습에는 이용하였다 [5]. Input layer에는 44개의 주제들로 노드들을 할당하였다. 각 Input 노드들에는 입력 문서의 $word_a$: $subject_s$.$word_b$들을 주제($subject_s$)별로 합한 값들이 입력된다. 이렇게 한 이유는 모든 $word_a$:$subject_s$.$word_b$을 Input vector로 사용할 경우에 비용이 너무 크기 때문이다. Hidden layer는 4개와 10개로 실험하였으며, Output layer는 노드 수를 1개로 유지하였다.



<그림 1> BPN 구성

3.3 기타 변수들

learning rate는 0.2, momentum값은 0.5로 설정하였다.

4. 학습 및 테스트

사람에 의해서 동물과 관련된 문서인지 아닌지 미리 평가해 놓은 관련문서 295개와 비관련문서 1000개가 학습과 테스트에 사용되었다.

4.1 Training Set

전체 문서 중에 220개의 관련문서와 750개의 비관련문서가 학습에 사용되었다. 학습은 비관련문서와 관련문서를 교대로 반복하면서 하였다. 비관련문서에 대해 수가 적은 관련문서를 계속 되돌려 가며 학습시켰다. 이렇게 총 1500개의 문서가 학습에 사용되었다. 출력 층의 목표값은 관련문서일 경우에는 1이, 비관련문서일 경우에는 0이 나오도록 학습시켰다.

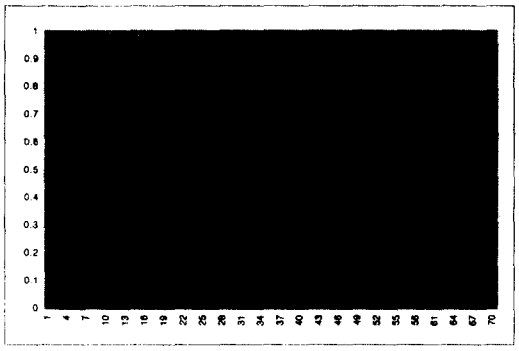
4.2 Test Set

관련문서 70개와 비관련문서 186개가 테스트에 이용되었다. 각 문서에 대해서 BPN의 출력값을 테스트 값으로 저장해 놓았다.

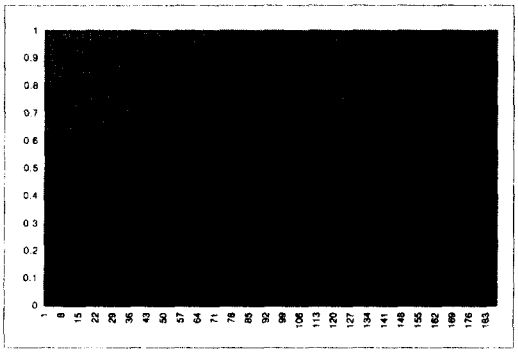
5. 결과

<결과 1>과 <결과 2>는 히든 노드의 수를 4개로 한 차례 학습한 결과이고, <결과 3>과 <결과 4>는 히든 노드의 수를 10개로 한 차례 학습한 결과이다. <결과 1>과 <결과 2>, <결과 3>과 <결과 4>를 묶어서 보면 관련문서와 비관련문서 간에 출력값이 큰 차이가 나는 것을 알 수 있다. 관련문서를 테스트한 <결과 1>과 <결과 3>을 보면, 대다수의 출력값이 0.5를 초과한다. 비관련문서를 테스트한 <결과 2>와 <결과 4>의 경우에는 출력값이 0.5 미만인 문서들이 대부분이다. 특히 히든 노드의 수가 10개인 <결과 4>의 출력값이 <결과 2>보다 작다는 것을 알 수 있다.

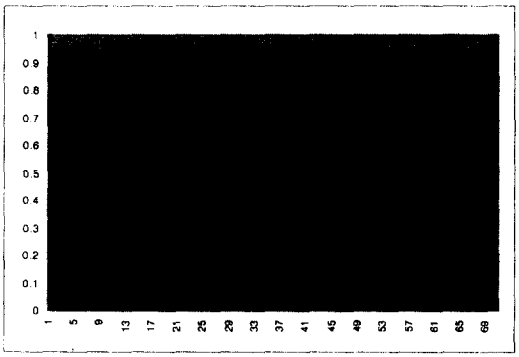
이 결과들은 한 차례 학습시킨 것이며, 학습을 더 진행시켰을 경우에는 관련문서와 비관련문서 사이의 차이가 커지는 결과를 보여주었다. 하지만, 몇몇 관련문서들의 출력값이 큰 폭의 감소를 보여주었는데, 그 이유는 Training Set에서의 비관련문서 수가 관련문서보다 많기 때문이다.



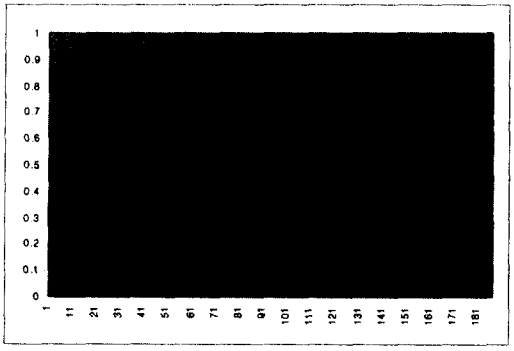
<결과 1> 관련문서, 히든 노드 수 4의 테스트 결과



<결과 2> 비관련문서, 히든 노드 수 4의 테스트 결과



<결과 3> 관련문서, 히든 노드 수 10의 테스트 결과



<결과 4> 비관련문서, 히든 노드 수 10의 테스트 결과

<결과 1>과 <결과 2>에서 출력값이 0.5 이상이면 관련문서로, 그 미만을 비관련문서로 평가하면 정확률은 70.31%(180/256)이다. <결과 3>과 <결과 4>에서는 출력값이 0.2 이상이면 관련문서로, 그 미만을 비관련 문서로 평가하면 정확률은 72.27%(185/256)이다. 정확률이 70%의 수준에 머물지만 관련문서를 비관련문서로 판단하여 제외시키는 것보다는, 관련문서는 가능한 모두 가지고 오면서 일부의 비관련문서를 관련문서로 가져오는 것이 이득이 더 크다. <결과 1>과 <결과 3>에서 관련문서의 출력값이 대부분 0.8이상의 높은 값을 보이고 있지만, 위와 같은 이유로 테스트의 정확률 계산에는 낮은 값을 적용하였다. 그리고, 관련문서로 평가된 비관련문서들이 사용자에게 실제로 서비스될 때에는 순위에서 밀리게 된다.

6. 결론 및 향후 연구

동물 정보를 담고 있는 웹 문서를 평가해 내는 한 방법으로 수치화된 단어와 단어의 관계를 주제라는 큰 범위로 묶어 학습시킨 BPN을 이용해 보았다. 이렇게 학습된 BPN에서 좋은 테스트 결과를 얻을 수 있다. 주제를 WordNet을 이용해 생성했기 때문에, 이 방법을 동물 주제 뿐만이 아니라 다른 주제에도 쉽게 적용해 볼 수 있을 것이다. 또한, 테스트 결과에 관련문서 출력값과 비관련문서의 출력값이 확연한 차이가 나는 것을 보면, 단어와 단어의 관계를 이용한 웹 문서 평가에 BPN만이 아니라 다른 방법을 이용해도 좋은 결과를 기대할 수 있을 것이다.

참고문헌

[1] 이용현, 홍익대학교 컴퓨터공학과, 정보통신망에서 지능형 정보 에이전트와 특정 영역에서의 구현. 1999.

[2] 김상모, 홍익대학교 컴퓨터공학과, 웹에서 특정영역 정보에이전트의 성능향상에 관한 연구. 2000.

[3] http://www.cogsci.princeton.edu/~wn/

[4] 이재윤, 김태수, 연세대학교 문헌정보학과, WordNet과 시소러스, 1999.

[5] James A. Freeman, David M. Skapura, "Backpropagation" Ch.3, Neural Networks Algorithms, Applications, and Programming Techniques. 1991