U-WIN을 이용한 WSD 기반의 문서 유사도 측정

심강섭[O1], 배영준[1], 옥철영[1], 최호섭[2]
[1]울산대학교 컴퓨터정보통신공학과
goodsoult@nate.com, young4862@ulsan.ac.kr, okcy@ulsan.ac.kr
한국과학기술정보연구원 정보기술개발단 정보시스템개발팀
hschoe@kisti.re.kr

Measurement of WSD based Document Similarity using U-WIN

Kang-Seop Shim[O1], Young-jun Bae[1], Cheol-Young Ock[1], Ho-Seop Choe[2]
Dept. of Computer Engineering and Information Technology, University of Ulsan
goodsoult@nate.com, young4862@ulsan.ac.kr, okcy@ulsan.ac.kr
Information System Development Team, Korean Institute of Science and Technology Information
hschoe@kisti.re.kr

요 약

이미 국외에서는 WordNet과 같은 의미적 언어자원을 활용한 문서 유사도 측정에 관한 많은 연구가 진행되고 있다. 그러나 국내에서는 아직 WordNet과 같은 언어자원이 부족하여, 이를 바탕으로 한 문서 유사도 측정 방법이나 그 결과를 활용하는 방법에 관한 연구가 미흡하다.

기존에 국내에서 사용된 문서 유사도 측정법들은 대부분 문서 내에 출현하는 어휘들의 의미에 기반하기보다는, 그 어휘들의 단순 매칭이나 빈도수를 이용한 가중치 측정법, 또는 가중치를 이용한 중요 어휘 추출방법들 이었다 . 이 때문에, 기존의 유사도 측정법들은 문서의 문맥정보를 포함하지 못하고, 어휘의 빈도를 구하기 위하여 대용량의 문서집합에 의존적이며, 또한 특정 개념(의미)을 다른 어휘로 표현하거나, 유사/관련 어휘가 사용된 유사 문서에 대한 처리가 미흡하였다.

본 논문에서는 이에 착안하여 한국어 어휘 의미망인 U-WIN과 문맥에 사용된 어휘들의 overlap 정보를 사용하여, 단순히 어휘에 기반하지 않고, 기본적인 문맥정보를 활용하며, 어휘의 의미에 기반을 둔 문서 유사도 측정법을 제안한다.

1. 서 론

유사문서 검색은 대상문서와 유사한 문서들의 유사도 순위를 매겨 사용자에게 검색결과로 제공한다. 이는 검색된 문서와 유사한 문서가 검색어와 관련이 있을 수 있고, 사용자가 검색하고자 하는 것에 대한 추가적인 정보를 제공할 수도 있기 때문에, 최근 정보 검색의 추세가 되고 있다[5].

이미 국외에서는 유사문서 검색이나 문서 유사도 측정에 관한 여러 연구가 진행되고 있으며, 그 중에는 WordNet과 같은 의미적 언어자원을 활용한 유사 문서 검색에 관한 많은 연구도 진행되고 있다.[1,2,3]. 국내에서는 아직 한국어로 구축된 의미적 언어자원이 부족하다. 따라서 문서 내에 출현하는 어휘의 의미에 기반하기보다는, 어휘의 단순 매칭이나 중요 어휘 선별, 또는 어휘의 빈도(TF-IDF)에 기반하여 가중치를 달리하는 문서 유사도 측정법이 주로 사용되고 있다[4,5]. 이러한 기존의 측정법들은 단순한 어휘 매칭을 이용함으로써 문맥정보를 활용하지 못하고, 또한 어휘의 빈도를 구하기 위해서는 대용량의 문서집합(*corpus*)이 필요하다는 단점을 가지고 있다. 더불어, 대상문서의 특정 개념(의미)이 여러 다른 어휘(동의어, 유의어)로 표현된 경우나, 그 개념과 관련된 부가적인 개념(관련어)들이 대상문서에는 나타나있지 않고 비교문서에만 나타나 있는 경우를 처리하는 데 미흡하다.

본 논문에서는 이러한 점에 착안하여, 문맥에 사용된 어휘들의 overlap을 이용하여 기본적인 문맥정보를 활용하고, 대상문서에 나타나지 않는 관련어휘(동의,유의,관련어)들의 처리를 위해, 한국어 의미망인 U-WIN을 이용하여 어휘의 의미에 기반을 둔 문서 유사도 측정법을 제안한다.

2. 관련 연구

문서의 유사도를 측정하는 방법은 크게 문서에서 추출된 어휘들만을 사용한 방법과, 그 어휘들의 가중치를 이용한 방법이 있다[5].

2.1 어휘들만을 이용한 유사도 측정 방법

문서에서 추출된 어휘들을 이용할 때는 대상문서와 비교문서의 전체 어휘 개수와 두 문서에 공통적으로 출현

한 어휘들의 개수를 사용하여 문서간의 유사도를 측정한다. 이 경우에 문서 유사도는 식 (1)과 (2)를 이용하여 계산할 수 있다.

$$Sim(T_i, T_j) = \frac{n_{ij}}{MIN(n_i, n_j)} \quad (1)$$

$$Sim(T_i, T_j) = \frac{n_{ij}}{MAX(n_i, n_j)} \quad (2)$$

n_{ij}는 대상문서와 비교문서에 공통으로 출현하는 어휘의 개수를 나타내며, $MIN(n_i, n_j)$과 $MAX(n_i, n_j)$는 각각 두 문서에서 추출된 어휘의 최소 개수와 최대 개수를 나타낸다.

2.2 어휘들의 가중치를 이용한 유사도 측정 방법

일반적으로 문서에서 추출된 어휘에 가중치를 할당하기 위해서는 TF(어휘 빈도수)-IDF(역문서 빈도수) 가중치 할당 방법을 사용한다. 이외에도 주제어 가중치 할당 방법 등을 사용해 가중치를 할당한다[12]. TF와 IDF는 각각 식(3)과 식(4)로 계산 할 수 있다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (3)$$

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (4)$$

식(3)에서 n_{ij}는 문서 d_j 안에서 어휘 t_i가 나타나는 개수이며, 분모는 문서 d_j 안의 모든 어휘의 개수를 나타낸다. 식(4)에서 $|D|$는 전체 문서 개수이며, $|\{d_j:t_i \in d_j\}|$는 단어 t_i가 등장하는 문서들의 개수를 나타낸다.

이렇게 어휘에 할당된 가중치를 이용한 대표적인 문서 유사도 측정법은 Cosine 유사도 측정법이며 식(5)와 같이 정의된다.

$$Sim(T_i, T_j) = \frac{\sum_{k=1}^{t}(w_{i,k} \times w_{j,k})}{\sqrt{\sum_{k=1}^{t} w_{i,k}^2} \times \sqrt{\sum_{k=1}^{t} w_{j,k}^2}} \quad (5)$$

$w_{i,k}$는 k번째 어휘의 가중치이며, t는 $|K_i \cup K_j|$ 이다.

3. U-WIN [User-Word Intelligent Network]

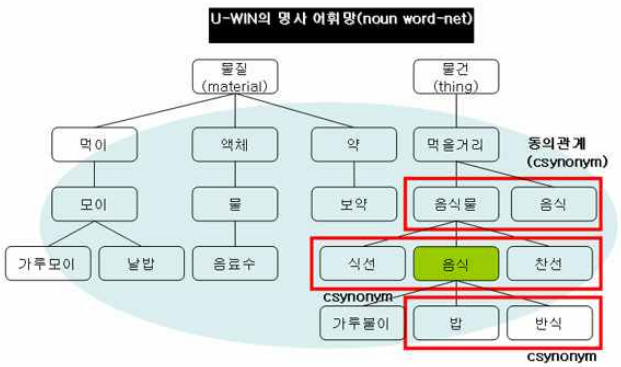
자연 언어의 어휘적 의미, 구문적 의미, 담화적 의미를 바탕으로 행위나 현상, 상태 등에 담긴 의미론적/개념론적 특성을 포함하고 있는 의미적 언어자원 구축에 대한 연구는 다양하게 이루어지고 있다. 국외에서는 WordNet, Euro WordNet, Cyc, HowNet, Lexical FreeNet, EDR 등이 대표적이며, 국내에서는 카이스트의 CoreNet, ETRI 의 어휘 개념망, 부산대의 KorLex 등이 대표적이라 할 수 있다[7].

KAIST CoreNet(코어넷)은 다국어 어휘 의미망으로서, 3,000여개의 개념이 한국어-중국어-일본어로 대응되어 동일한 개념체계를 공유한다.

ETRI 명사 개념망은 한국어 명사 어휘들을 의미관계로 연결시켜 놓은 어휘 데이터베이스로서, 깊이 12레벨로 구성되어 있으며, 약 25여 만 고유명사들이 Instance_Of 관계로 연결되어 있다.

U-WIN(User-Word Intelligent Network)은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적/개념적 네트워크를 형성한 온톨로지적 의미망이라 할 수 있다.

U-WIN은 현재 30만 여 어휘가 구축된 상태이다. U-WIN의 구축 대상은 한국어 어휘 전체(모든 품사 및 언어 단위)로서, 핵심적 대상은 명사, 동사, 형용사이며, 부수적 대상은 부사, 관형사, 대명사, 감탄사, 조사, 수사, 의존명사 등이며, 북한어, 방언, 옛말, 전문용어, 고유명사, 어근, 어미 등 한국어 어휘 전체를 대상으로 연구 개발 중이다. [그림 1]은 구축 사례의 일부 모습이다.



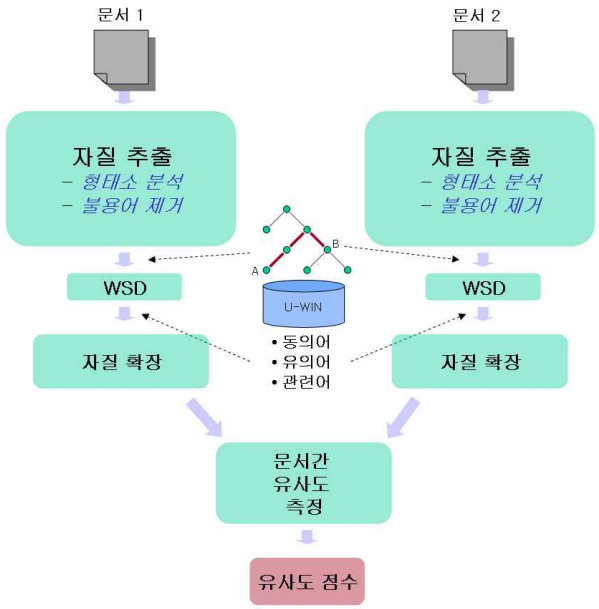
<그림 1> U-WIN의 구축 사례

U-WIN은 한국어 정보 처리를 비롯하여 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 대규모 어휘 지식 베이스를 목표로 하고 있다. 현재 온톨로지 기반 의미적 주석(ontology-based semantic annotation)과 유사한 단어 중의성 해소(word sense disambiguation)와 의미 태깅(semantic tagging) 기술에

활용되고 있으며, 이 외에도 복합명사 자동 생성, 전문 분야별 개념체계 자동 생성, 정보검색에서의 질의확장, 어휘 학습 시스템 등 다양한 기술에서 활용되고 있다.

3. 제안한 문서 유사도 측정법

본 논문에서는 대상문서와 유사한 문서를 검색하기 위해서, [그림 2]와 같이 자질추출(형태소 분석, 불용어 제거), WSD, 자질확장, 문서간 유사도 측정의 과정을 단계적으로 처리한다.



<그림 2> 제안된 문서 유사도 측정의 전체 처리 과정

3.1 자질 추출 (Feature Extraction)

본 논문에서는 문서간의 유사도를 측정하기 위해 주어진 문서에 대한 형태소 분석과 불용어 제거로 자질을 추출하였다. [그림 3]은 주어진 문서에서 자질을 추출한 모습이다. "#IDX"는 추출된 자질을 나타낸다.

본 논문은 초전도 및 극저온 케이블의 냉매와 전기절연으로 사용되고 있는 액체질소의 연면방전특성에 관하여 연구하였다. 기포효과를 고려하여 대기압하에서의 정극성 및 부극성 직류고전압을 사용하였고, 연면방전시 방사되는 방사전자파의 스펙트럼 분포 특성을 관측하였다.

자질 추출

#IDX=초전도&극저온&케이블&냉매&전기&절연&액체&질소&연면&방전&연구&기포&대기압&정극성&부극성&직류&고전압&연면방&전시&연면방전&연면&방전&방사&방사&전자파&스펙트럼&분포&관측

<그림3> 자질 추출

3.2 WSD (Word Sense Disambiguation)

WSD(Word Sense Disambiguation)는 문서 내에 출현하는 자질들의 의미를 결정하는 과정이다. 본 논문에서는 U-WIN의 계층 구조를 이용하여 Path Length 기반 의미간 유사도 측정 방법과 Patwardhan et al.[11]의 알고리즘을 사용하여 WSD를 하였으며, 그 대상은 U-WIN에 나타나 있는 자질들이다.

3.2.1 Path Length 기반 측정 방법

Path Length를 기반으로 한 의미 유사도 측정방법은 계층 구조상에서의 개념간 최단 경로 수를 계산하거나 개념의 깊이, 관계 종류 등을 고려할 수 있다. 대표적인 Path Length 기반 개념간 유사도 측정 방법에는 Rada, et. al[7]., Leacock and Chodorow[8], Wu and Palmer[9] 등이 있다.

Rada, et. al[7]은 두 개념간 최단 경로를 기반으로 유사도를 측정하였으며, Leacock and Chodorow[8]는 Rada, et. al[7]의 방법에 계층구조의 최대 깊이를 함께 고려하였다.

$$Sim_{lch}(c_1, c_2) = -\log(\frac{dist_{node}(c_1, c_2)}{2 \cdot D}) \qquad (6)$$

식(6)에서 *dist*는 두 개념 사이의 최단 경로이고, *D*는 계층 구조의 최대 깊이이다.

그리고 Wu and Palmer[9]은 계층 구조에서 개념의 깊이에 기반하여 개념간 유사도를 측정하였다.

$$Sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \qquad (7)$$

수식(7)에서 *depth*는 계층 구조의 가장 상위인 *root*로부터 개념 c 까지의 거리를 뜻하며, *lcs*(least common subsumer)는 계층 구조에서 개념 c_1과 c_2를 모두 포함하는 가장 하위의 개념을 뜻한다.

본 논문에서는, 위의 세 가지 유사도 측정 방법 중에서 U-WIN에 적용했을 때, 가장 좋은 성능을 보인 Wu and Palmer[9]를 개념 간 유사도 측정 방법으로 선정하였다.

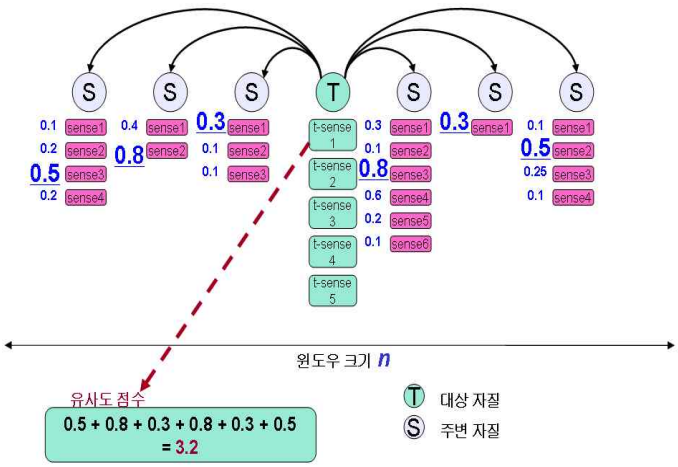
3.2.2 WSD 알고리즘

Patwardhan et al[10]는 문서의 시작위치에서 오른쪽으로 진행하면서 자질의 의미(sense)를 결정한다. 각각의 자질들은 한 개 이상의 의미를 가지며, 그 중 하나를 자질의 의미로 결정한다. 의미가 결정될 자질을 대상자질이라고 하며, 대상자질을 제외한 크기 n 윈도우안의 자질들을 주변자질이라고 한다.

Patwardhan et al[10]은 대상자질의 각 의미 즉, 대상의미마다 점수를 할당하여 가장 높은 점수를 가진 대상의미를 그 의미로 결정한다. 점수할당을 위하여 우선, 대상의미와 가장 의미 유사도가 높은 주변자질의 의미들을

찾는다. 그리고 그렇게 찾아진 주변의미들의 유사도 합을 대상의미에 할당하게 된다. 전체 과정은 다음과 같다.

1 대상 자질이 중간에 위치하도록 크기 *n*의 윈도우를 선택한다.
2 윈도우 안의 각 자질들의 후보의미를 사전에서 추출한다.
3 대상 자질의 각 의미들에 대해서 :
  3.1 대상 자질의 각 의미들과 주변 자질의 의미들과의 의미 유사도를 측정한다.
  3.2 각 주변 자질의 의미들이 가지는 유사도들 중, 가장 큰 유사도들의 합계를 구한다.
  3.3 합계를 해당 대상 의미에 할당한다.
4 가장 높은 점수를 가지는 대상 의미가 대상 자질의 의미로 결정된다.



<그림 4> WSD 알고리즘의 예

또한 Wu and Palmer[9]을 사용한 WSD 알고리즘은 식(8)으로 나타낼 수 있으며, 총 계산 횟수는 수식(9)와 같다[11].

$$\arg\max_{i=1}^{m_t} \sum_{j=1, j\neq t}^{n} \max_{k=1}^{m_j} Sim_{wup}(s_{ti}, s_{jk}) \qquad (8)$$

$$\#Computations = \sum_{i=1, i\neq t}^{n} m_t \cdot m_i \qquad (9)$$

식(8)에서 n은 윈도우 크기이며, n개의 자질들 $w_1, w_2, \ldots, w_n$ 에 대해서, $w_t (1 \le t \le n)$은 대상자질 즉, 의미가 결정될 대상이며, w_t가 m_t개의 의미를 가진다고 할 때, 각각의 의미(*sense*)들은 s_{t1}, s_{t2}, …, s_{tmt}으로 나타낸다. 예를 들어, 대상자질의 각 의미들은 {s_{t1}, s_{t2}, … , s_{tmt}}와 같이 나타낼 수 있다.

[그림 5]는 WSD가 된 자질들의 모습이며 "스펙트럼/23235301"과 같은 형태를 취한다. "23235301"은 U-WIN내에서 할당된 의미의 ID이다.

3.3 자질 확장 (Feature Extension)

WSD된 문서는 의미 ID를 이용해서 U-WIN으로부터 각 자질들의 동의어, 유의어, 관련어를 추출하여 확장을 한다. [그림 6]은 자질확장을 한 모습이다. 각 자질들의 동의어, 유의어는 "고전압:고압:고압전기"와 같은 형식으로 "#IDX"에 추가하였으며, 각 자질들의 관련자질들은 "#RDW"에 추가하였다.

U-WIN으로부터 관련어들을 추출할 때는, 각 자질들과 해당 관련어들의 직접적인 상관도를 고려하여 그 *depth*를 제한하였으며, 본 논문에서는 그 *depth*를 3으로

<그림 5>WSD

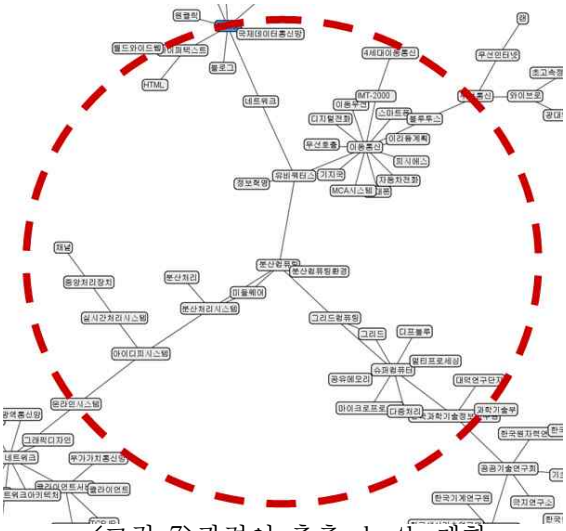
<그림 6>자질 확장

로 정하였다. [그림 7]은 U-WIN에 구축된 특정 개념과 관련어들에 대한 예시이며, 점선은 *depth* 제한을 나타낸다.

3.3 문서간 유사도 측정(Similarity Measure)

입력된 두 문서간의 유사도 계산은 기본적으로 문서내 출현하는 자질들간의 공기정보를 이용하며, "#IDX"와 "#RDW"에 속한 자질의 점수 계산 방법은 서로 다르다.

"#IDX"에 속한 자질들에 대해서는 기본적인 문맥정보를 반영하기 위해서, 뜻풀이를 이용한 의미간 유사도 측



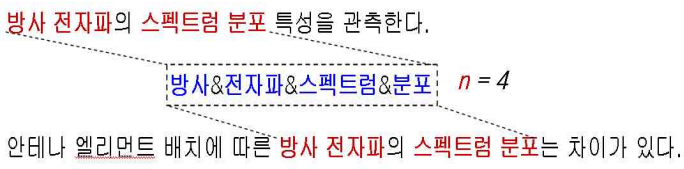
<그림 7>관련어 추출 depth 제한

정법 중의 하나인, The Adapted Lesk Algorithm[11]을 사용하여 점수를 할당한다.

The Adapted Lesk Algorithm[11]은 두 문서를 비교할 때, overlap의 개념을 사용한다. overlap은 두 문서에서 공통적으로 출현하는, 순차적이며 연속적인 가장 긴 패턴을 의미한다. The Adapted Lesk Algorithm[11]은 overlap의 길이가 길수록 문서에서 나타날 확률이 낮다는점에 기반하여, 그 길이에 따라 가중치를 주며, $(length\ of\ overlap)^2$으로 점수를 계산한다. 예를 들어, 길이가 1인 4개의 overlap들은 4점으로 계산하지만, 길이가 4인 1개의 overlap은 16점으로 계산한다. 그림 8에서와 같이 4개의 자질 {방사, 전자파, 스펙트럼, 분포}가 연속적으로 나타날 경우 16점으로 계산한다. 따라서 "#IDX"에 대한 두 문서간의 점수는 식(10)이 된다.

$$score_{idx} = \sum_{i}^{\#overlaps} length^2(overlap_i) \qquad (10)$$

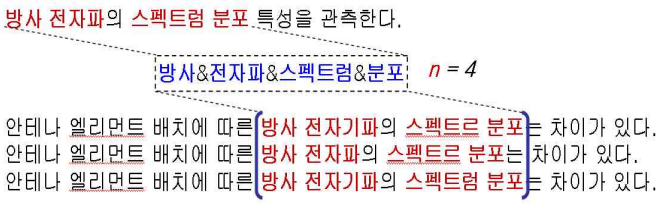
*length*는 overlap의 길이이며, *#overlaps*는 overlap의 총 개수이다.



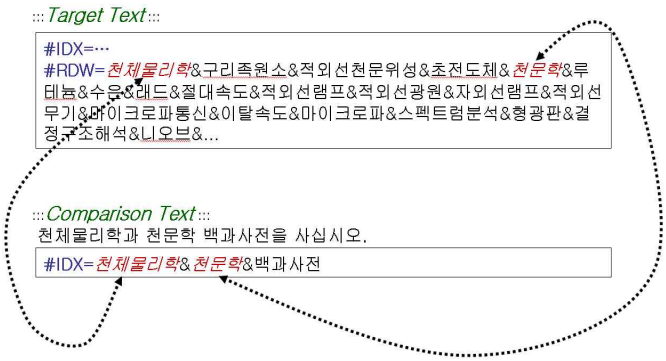
<그림 8> overlap을 이용한 점수 계산 - 1

또한, 본 논문에서는 같은 개념을 사람마다 다른 어휘로 표현할 수 있다는 점에 착안하여 동의어, 유의어로 문맥이 대체된 문서에도 같은 점수를 할당한다. 예를 들어, {전자파, 전자기파}, {스펙트럼, 스펙트르}는 각각 동의어들의 집합이며, [그림 9]와 같은 비교가 가능하다.

"#RDW"는 관련 자질의 단순 집합으로서 문맥을 대체하지 못하기 때문에 각 자질 당 매칭 개수의 합을 점수로 할당하며, 대상문서에 나타난 자질들의 관련 자질이 비교문서의 본문에 많이 나타날수록 두 문서의 유사도가 높을 수 있다는 점에 착안하여 비교문서의 본문인 "#IDX"와 비교한다. 예를 들어, [그림 10]과 같이 대상문서의 "#RDW"와 비교문서의 " #IDX"에 2개의 자질이 동시에 출현하면 2점을 할당한다.



<그림 9> overlap을 이용한 점수 계산 - 2



<그림 10> RDW의 점수 계산

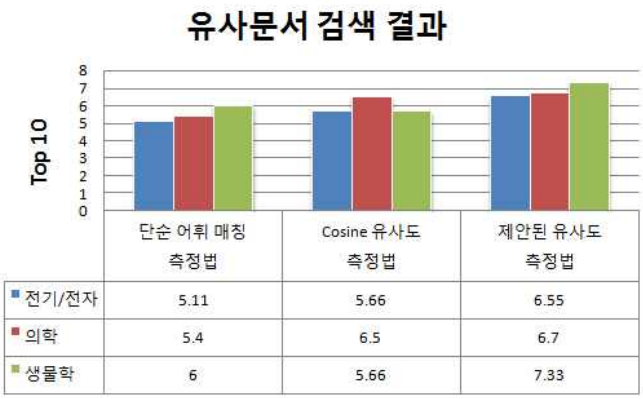
최종적으로 "#IDX"의 유사도 점수와 " #RDW"의 유사도 점수의 합을 두 문서간 유사도 점수로 한다.

4. 실험 및 평가

본 논문은 제안한 문서 유사도 측정방법의 신뢰성을 입증하기 위하여, 전기/전자, 의학, 생물학 분야 한국어 논문 각각 2,000개의 요약을 대상으로 단순 어휘 매칭, Cosine 유사도 측정법, 제안한 유사도 측정법을 이용하여 유사문서 검색 실험을 수행하였다. 유사문서 검색을 수행하기 위하여 각 분야의 전체 문서 중, 대상문서로 사용할 각 100개의 문서를 무작위로 선정하며, 대상 문서에 대해 다음과 같은 과정을 단계적으로 처리한다.

1. 대상 문서에 대한 비교문서들의 유사도 점수를 각각 세 개의 유사도 측정법을 이용하여 할당한다.
2. 점수가 할당된 비교문서를 내림차순으로 정렬한다.
3. 각 측정법 별로 정렬된 비교문서에서, 최상위 10개의 문서 중에 몇 개의 정답 문서를 포함하고 있는지 개수를 샌다.

정답 문서는 동일한 주제, 키워드, 방법론을 다룬 문서라고 정의하였다. [그림 11]은 실험 결과를 도표로 나타낸 것으로, 제안한 방법이 유사문서 검색의 정확도 면에서 단순 어휘 매칭이나 Cosine 유사도 측정법보다 우수함을 알 수 있다.



<그림 11>실험 결과

5. 결론

유사문서를 검색하는 데 있어서, 문서 내 출현하는 어휘의 의미를 파악하는 것이 중요하다. 그럼에도 불구하고 국내에는 의미적 언어자원이 부족하여, 어휘에 기반한 유사문서 검색 연구들이 주로 진행되고 있다.

본 논문에서는 문서 내 출현하는 어휘에 기반한 문서간 유사도 측정 방법에서 탈피하여, 한국어 의미망인 U-WIN을 이용한 의미 기반의 문서간 유사도 측정 방법을 제안하였고, 각 2,000개의 전기/전자, 의학, 생물학 분야 문서를 대상으로 실험하여 실험결과와 같이, 일반적으로 많이 쓰이는 단순 어휘 매칭이나 Cosine 유사도 측정법보다 더 우수한 성능을 보였다.

향후 과제로서, 좀 더 다양한 영역의 문서에 대한 실험과 WSD 정확도 증가 그리고 단순 의미에 기반한 측정법이 아니라, 의미의 중요도에 기반한 문서 유사도 측정법으로의 개선이 필요하다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.(IITA-2008-(C1090-0801-0039))

참고 문헌

[1] P.Selvi and Dr.N.P.Gopalan, “SEMANTIC TEXT SIMILARITY COMPUTATION USING MULTIPLE INFORMATION SOURCES“, International Journal of Computer Science and Network Security, VOL.7, No.12, December 2007.

[2] Abhinay Pandya and Pushpak Bhattacharyya, “Text Similarity Measurement Using Concept Representation of Texts“ LNCS 2776, pp.678-683, 2005.

[3] Thorsten Brants and Reinhard Stolle., “Finding Similar Texts in Text Collections“, InProceedings of the 3rd International Conference on Language Resources and Evaluation(LREC-2002), Workshop on Using Semantics for Information Retrieval and Filtering, Las Palmas, Spain, 2002.

[4] 장성호, 강승식, “용어 선별 기법에 의한 유사 문서 판별 시스템”, 한국정보과학회 봄 학술발표논문집, Vol.30, No.1, 2003. 4.

[5] 김혜숙, 박상철, 김수형, “단어/단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정”, 한국정보과학회 논문지 : 소프트웨어 및 응용, Vol.31, No.12, 2004. 12.

[6] 최호섭, 임지희, 옥철영, “A Case Study of Ontology Construction“, 한국정보과학회 논문지, Vol.24, No.4, pp.31-44, 2006.

[7] R. Rada, H, Mili, E. Bicknell, M. Blettner, “Development and application of a metric on semantic nets, IEEE Transactions on Systems“, Man and Cybernetics, 19(1), pp.17-30, 1989.

[8] C. Leacock, M. Chodorow, “Combining local context and WordNet similarity for word sense idntification“, in: C. Fellbaum (Ed.), WordNet: An electronic lexical database MIT Press, pp. 265-283, 1998.

[9] Wu, Z., Palmer, “Verb semantics and lexical selection“, 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, LasCruces, New Mexico, 1994.

[10] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen, “Using measures of semantic relatedness for word sense disambiguation.“ In Proceedings of the Fourth Internatinoal Conference on Intelligent Text Processing and Computational Linguistics, pp.241-257, Mexico City, February 2003.

[11] Pedersen, Ted and Banerjee, Satanjeev, and Patwardhan, Siddharth, “Maximizing Semantic Relatedness to Perform Word Sense Disambiguation“, University of Minnesota Supercomputing Institute Research Report UMSI 2005/25 March 2005.

[12] 강승식, 이하규, 손소현, 홍기채, 문병주, “조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법”, 한국정보과학회 가을 학술발표논문집, Vol.28, No.2, pp.196-198, 2001.

[13] 전문용어공학센터[KORTERM], 『다국어 어휘망』 총 3권, KAIST Press, 2005.

[14] 최호섭, “한국어 명사 개념망 구축-경제용어를 중심으로”, ETRI 지식정보검색연구팀 경제개념망 구축 결과보고서, 2001.