문서분류를 위한 의미적 주제선정방법

고광섭* · 황명권** · 김판구*** · 이창훈****

Semantic Topic Selection Method of Document for Classification

Kwang Sup Ko* · Myung Gwon Hwang** · Pan Koo Kim*** · Chang Hoon Lee****

요 약

웹은 전세계 규모의 네트워크로써 문자, 화상, 음성 등의 미디어 정보들을 페이지 단위로 관리되며, 링크를 이용하여 분산된 정보들을 연결하고 있다. 이러한 웹의 지속적인 발전으로 무수한 정보들을 축적하고 있으며, 그 중 텍스트로 구성된 문서들이 주를 이룬다. 사용자는 이렇게 많은 정보들 중에서 자신이 원하는 특정 정보를 찾기 위해 웹을 사용한다. 그래서 웹은 사용자 요구에 적합한 정보를 검색해 주기 위해 계속적인 시도와 많은 연구들로 발전되고 있다. 확률을 이용한 방법, 통계적인 기법을 이용한 방법, 벡터 유사도를 이용한 방법, 베이지안 자동문서 분류 방법 등 기존의 방법들은 문서의 의미적인 주제나 특징을 정확하게 처리할 수 없어 사용자는 재검색을 해야 하는 문제점을 갖는다. 특히, 국내 문서분류를 위한 연구는 많이 이루어지지 않아 검색에 더욱 어렵다. 이러한 문제점을 보완하기 위해 본 논문에서는 국내문서의 효율적이고 의미적인 분류를 위해 출현 개념의 TF(Term Frequency)와 주변 개념들과의 관계된 정도(RV : Relation Value)를 추출한다. 그리고 추출된 키워드들을 국내 어휘 사전인 U-WIN에 매핑하여 문서의 주제를 선택하고 본문에서 제시하는 분류방법에 의해 웹 문서를 분류한다. 이는 문서 내 개념들의 관계를 이용하여 문서의 주제를 선정하고 문서의 의미적인 분류를 가능하게 한다.

ABSTRACT

The web as global network, includes text document, video, sound, etc and connects each distributed information using link. Through development of web, it accumulates abundant information and the main is text based documents. Most of user use the web to retrieve information what they want. So, numerous researches have progressed to retrieve the text documents using the many methods, such as probability, statistics, vector similarity, Bayesian, and so on. These researches however, could not consider both the subject and the semantics of documents. As a result user have to find by their hand again. Especially, it is more hard to find the korean document because the researches of korean document classification is insufficient. So, to overcome the previous problems, we propose the korean document classification method for semantic retrieval. This method firstly, extracts TF value and RV value of concepts that is included in document, and maps into U-WIN that is korean vocabulary dictionary to select the topic of document. This method is possible to classify the document semantically and showed the efficiency through experiment.

키워드

문서분류, TF, 클러스터링, 주제선정

* 건국대학교 컴퓨터공학과 박사과정
** 교신저자 : 조선대학교 컴퓨터공학부 박사과정
*** 교신저자 : 조선대학교 컴퓨터공학부 교수
**** 교신저자 : 건국대학교 컴퓨터공학과 교수

접수일자 : 2006. 12. 5

Ⅰ. 서 론

웹의 성장으로 웹에는 무수한 문서들이 존재한다. 이러한 문서들을 수동으로 분류하기란 불가능하다. 이에 많은 연구자들이 의미적으로 웹 문서들을 처리하여 분류하기 위한 연구에 초점을 맞추고 있다.

본 논문에서는 국내문서의 분류를 위한 의미적 주제 선정에 초점을 맞추고, 효율적이고 의미적인 문서 분류를 위해 출현 개념의 TF(Term Frequency)와 주변 개념들과의 관계된 정도(RV : Relation Value)를 추출하고, 국내 어휘 사전인 U-WIN(UOU-Word Intelligent Network)을 이용한다. 문서 분류를 위해 본 논문에서 제시하는 방법은 다음과 같은 과정으로 구성되어 있다. 가장 먼저, 개념의 TF와 RV를 이용하여 문서의 키워드들을 추출한다. 추출된 키워드들을 U-WIN에 매핑을 시킨 후, 본 연구에서 제안하는 문서분류 방법에 의해 웹 문서를 적절하게 분류한다. 본 연구에서 제안하는 문서분류 방법은 3가지 단계로 구성되어 있으며, 실험을 통하여 정확하고 의미적인 문서분류가 가능함을 보였다.

본 논문은 2장에서 본 연구에 필요한 요소들을 상세히 설명하고, 3장에서 본 논문의 핵심인 문서 분류 방법을 각 모듈별로 기술한다. 4장에서 실험을 통해 본 연구에서 제안한 방법을 평가하고, 5장에서 결론 및 향후 연구방향을 제시한다.

Ⅱ. 관련연구

웹 문서들을 분류하기 위해 도메인 온톨로지와 프린스턴 대학의 워드넷(WordNet)을 이용한 연구가 많이 존재하였다[1][2][3][4][5][6]. 이러한 방법들은 문서의 키워드, 타이틀, TF등의 방법을 통해 핵심이 되는 개념들을 온톨로지 개념에 매핑을 하여 문서들의 특징들을 추출하였다[16]. 이들 연구들은 주로 영어로 작성된 문서 분류에 사용되며, 한글로 작성된 국내문서의 분류를 위해서는 한국어 사전을 이용해야 한다. 본 논문에서는 문서 분류를 위해 U-WIN 계층구조를 이용하여 접근하였으며, 문서내의 특정 개념의 중요도와 U-WIN에 정의된 관계를 최대한 반영하기 위해 문서내 개념들의 TF값과 개념 사이의 관계한 정도까지 고려하였다.

본 장에서는 본 논문에서 사용하고 있는 단어들의 빈도를 추출하는 TF방법과 문서의 의미적인 분류를 위해 필요한 U-WIN에 대해 설명한다.

2.1. TF (Term Frequency)

TF는 문서내에 포함된 개념들의 빈도를 나타낸다. 이는 특정 문서에 포함된 개념의 중요성을 측정하기 위한 척도로 사용된다[10]. 일반적으로, TF는 idf(inverse Document Frequency)와 결합하여 그룹내의 문서들 집단에서 특정 키워드에 맞는 문서검색 또는 마이닝에 사용된다. 하지만, 본 연구에서는 단순히 웹 문서에 포함된 키워드들의 빈도를 구하고 U-WIN에 매핑하기 위해 TF만을 이용하였다. 문서내의 개념들의 TF를 구하기 위한 식은 식 (1)과 같다.

$$tf = \frac{n_i}{\sum_k n_k} \tag{1}$$

식 (1)에서 $\sum_k n_k$는 문서 내에 포함된 개념들의 총합을 의미하고, n_i는 특정 개념의 출현 횟수를 나타낸다. 본 논문에서 TF는 문서 내에서 특정 개념이 얼마나 중요한지를 표현하는 척도로 사용된다.

2.2. U-WIN (UOU-Word Intelligent Network)

U-WIN은 울산대학교 한국어처리연구실의 옥철영 교수 연구팀에서 구축하고 있으며, 버전 1.0에서 동의어를 포함하여 18만개 이상의 어휘를 정의하고 있다. U-WIN은 한국어정보처리를 비롯한 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 응용될 수 있는 어휘 데이터베이스이며, '인간이 가지는 여러 관념 속에 공통적인 속성을 기반으로, 인간의 보편적인 인지 체계와 개념 관계를 파악하여 이것을 표현한 언어를 대상으로 한 형식적이고 명세적인 어휘 네트워크'라고 [11]에서 정의하고 있다. U-WIN은 DAT 파일에 개념들을 상위어, 하위어, 동의어와 의미를 중심으로 정의하고 있고, 몇 가지 기호들을 이용하여 개념들 사이의 관계를 정의하고 있다. 표 1은 개념들 사이의 관계를 나타내는 기호와 그 의미를 보이고 있다.

표 1. U-WIN의 개념관계 정의 기호와 의미
Table. 1 Symbol and Meaning of U-WIN

기호	의미
@	상위어
#	하위어
$	동의어

본 논문에서는 U-WIN을 이용하여 웹 문서에서 명사개념들을 추출하고, 표 1의 기호를 이용하여 개념들 사이의 관계를 파악하여, 마지막으로 상하관계의 기본 속성을 이용하여 문서의 핵심을 파악하고 분류를 시도하였다.

Ⅲ. 문서 분류 방법

본 논문에서 제시하는 문서분류방법은 다음과 같은 과정으로 구성되어있다.

가장 먼저, U-WIN을 이용하여 문서에 포함되어 있는 명사개념들을 추출하여 이들 각각의 TF값을 측정한다. 그리고 표 1의 기호를 이용하여 추출된 명사개념들 사이의 관계를 파악하여 핵심 키워드 집합을 파악한다. 핵심 키워드 집합의 개념들을 이용하여 문서의 주제를 선정하기 위하여, 각 개념들의 TF값과 개념들 사이에서 관계한 횟수(RV)를 이용하여 개념가중치(CW : Concept Weight)를 구한다. 이 과정에서 TF와 RV 두 가지를 모두 고려한 이유는 문서에서 중요한 개념일수록 출현 및 주변개념과의 관계된 횟수가 많다는 가정과 실험결과의 증명에 의한 것이다. 각 개념가중치 CW를 이용하여 문서를 분류하기 위해 표 2와같이 3가지 과정을 거친다.

표 2. 문서 분류 과정
Table. 2 Process of Document Classification

과정	조건	예외
1	특정 개념의 CW값이 다른 개념들의 CW값의 합보다 클 때, 특정 개념을 문서의 주제로 판단	만족하지 않을 경우 과정 2 진행
2	하위에 있는 개념으로 상위개념의 CW값을 상속하여 과정 1 적용	만족하지 않을 경우 과정 3 진행
3	상위에 있는 개념으로 하위 개념의 CW값들을 역상속하여 과정 1 적용	

3.1. 핵심 키워드 그룹 파악

핵심 키워드들의 그룹을 파악하는 것이 문서 분류 과정에서 가장 중요한 역할을 수행한다. 본 연구는 문서에 포함된 개념들 중에서 문서의 주제 선정에 주요한 역할을 하는 개념들은 출현 횟수와 개념들 사이의 관계가 많을 것이라는 가정 하에 진행되었다. 핵심이 되는 키워드들을 선택하기 위해 문서에 포함된 개념들의 TF와 RV를 파악하는데, 이는 의미적인 키워드 선정에 유용하고 문서의 주제와 가장 관계 깊은 개념들을 추출하는 것이 가능하다.

개념들의 TF와 RV를 이용하여 핵심 키워드 그룹을 파악하는 과정은 아래와 같다.

(과정 1) 문서에 포함된 명사개념을 추출하고 각 개념의 TF를 측정한다. 명사개념의 추출을 위해 형태소 분석 기법을 사용하여 구문태깅을 하고 명사만을 추출하였다. 이는 한국어 문서를 표현하고 이해하는데 효과적이다[12].

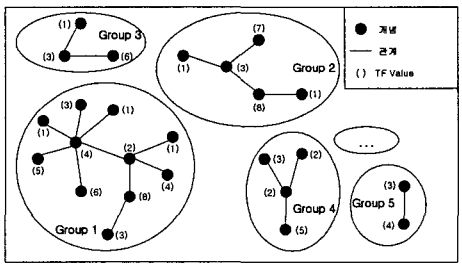

그림 1. 개념 그룹핑
Fig. 1. Concepts Grouping

(과정 2) 추출된 모든 개념을 U-WIN의 개념들과 매핑하고, 그림 1과 같이 U-WIN에서 정의하고 있는 관계, 즉 상/하위관계, 동의관계, 동위관계(형제계층)까지 각 개념들을 그룹핑한다. 그림 1의 개념 그룹핑 과정에서, Group 1에서 $CW_{1i}(4\times6)$은 TF 값 4와 RV 값 6을 의미한다.

(과정 3) 과정 2에서 얻어진 두 값을 이용하여, 문서의 키워드들의 그룹 선정을 위한 식 (2)를 적용한다.

$$Max(cwG_j) = \sum TF_i \times RV_i \qquad (2)$$

식 (2)에서 cwG_j는 각 그룹 가중치를 나타내고, TF_i는 특정 개념의 TF 값이며 RV_i는 문서내에 관계된 개념들의 개수를 나타낸다.

(과정 4) 식 (2)를 통해 계산된 각 그룹의 cwG_j 값이 가장 높은 그룹을 키워드 그룹으로 선정한다.

표 3은 그림 1을 이용하여 각 그룹의 cwG_j 값을 측정한 결과이다. 표 3의 결과에 의해 Group 1의 값이 72로 가장 높아 키워드 그룹으로 선정된다.

표 3. 그림 1과 식 (2)를 이용한 키워드 그룹 선정
Table. 3 Keyword Group Selection using Fig. 1 and Formula (2)

Group	Concept Weight $\sum CW_i(TF_i \times RV_i)$	cwG_j
1	4×6+1×1+3×1+1×1+5×1+6×1+2×4+1×1 +4×1+8×2+3×1	72
2	3×3+8×2+1×1+7×1+1×1	34
3	3×2+6×1+1×1	13
4	2×3+5×1+3×1+2×1	16
5	3×1+4×1	7

핵심 키워드를 추출하는 본 과정에서 얻어진 키워드 그룹과 각 개념의 CW값들은 문서 분류를 위한 주제를 선정하는 과정에서 다시 사용된다.

3.2. 문서 분류를 위한 주제 선정

문서의 주제를 선정하기 위해, 앞의 과정에서 추출된 키워드 그룹, 각 개념의 CW값과 U-WIN에 정의된 개념들의 계층구조를 다시 이용한다. 추출된 키워드 그룹의 개념들을 U-WIN의 개념들과 매핑을 시킨 후, 문서의 주제를 선정하기 위해 3가지 과정을 제안한다. 3가지 과정을 단계별로 수행함으로써 문서의 주제를 선택하고 분류할 수 있음을 보인다.

3.2.1 주제 선정 : 과정 1

과정 1에서는 키워드 그룹에서 특정 개념의 CW값이 다른 키워드들의 CW값의 합보다 크다면, 개념들 사이의 관계를 고려하지 않고 CW값에 의해 주제를 선정하는 방법이다. 이 방법은 특정 개념을 주제로 문서가 기술될 때, 주제가 되는 개념을 중심으로 다수의 주변 개념이 기술이 되고, 출현 횟수가 많다는 것에 근거에 의하고 있다. 과정 1을 수행하는 절차는 다음과 같다.

(1) 키워드 그룹 내에서 특정 개념을 제외한 나머지 개념들의 집합 CS_i를 구한다.

$$CS_i = \{C_1, C_2, C_3, ..., C_n\} \cap \{C_i\}^C$$

(2) 문서의 주제 개념을 선정하기 위해 식 (3)을 적용한다. 식 (3)은 추출된 키워드 그룹에서 특정 개념의 CW가 CS_i 내의 모든 개념의 CW값보다 크다면 주제로 선정하기 위함이다.

$$if(CW(C_i) \geq \sum CW(CS_j)) \quad (3)$$
$$then(Topic = C_i)$$

(3) 식 (3)을 만족하는 특정 개념이 존재하면 그 개념을 문서의 주제로 선정하고, 어떤 개념도 만족하지 않는다면 과정 2로 넘어간다.

그림 2와 표 4는 본 과정을 이용하여 문서의 주제를 선정하는 방법을 보이고 있다.

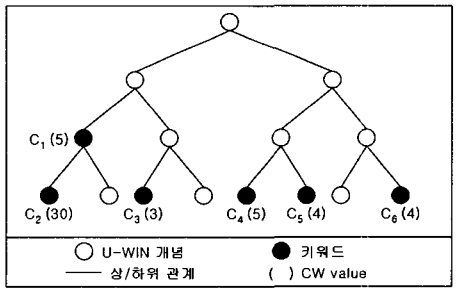

그림 2. 과정 1을 이용한 주제 선정 방법
Fig. 2. Topic Selection Method using Process 1

표 4. 과정 1을 이용한 주제 선정
Table. 4 Topic Selection using Process 1

개념	$CW(C_i)$	$\sum(CW(CS_i))$	결과
C_1	5	46	X
C_2	30	21	주제선정
C_3	3	48	X
C_4	5	46	X
C_5	4	47	X
C_6	4	47	X

표 4는 과정 1에 의한 주제 선정 결과를 보이고 있다. 과정 1에 의해 추출된 키워드 그룹에서 C_2가 문서의 주제로 선정되고, 과정 1에 의해 만족되는 결과가 존재하지 않는다면, 과정 2를 수행하게 된다.

3.2.2 주제 선정 : 과정 2

과정 1을 통하여 주제가 선택되지 않은 것은, 추출된 키워드 그룹에 포함된 개념들의 분포와 관계가 하나의 핵심 주제 개념에 치우치지 않은 것이다. 이는 두 가지 경우로 나누어 볼 수 있다. 첫 번째는 상위개념의 내용을 기반으로 하위개념을 설명하는 경우와, 여러 하위개념들을 설명하여 상위 개념을 부각시키려는 경우이다. 과정 2에서는 전자의 경우의 문서를 분석하는 방법으로 상/하위 개념 계층 구조에서 하위개념이 상위개념의 내용을 상속받는 전이적 속성을 전제로 주제 선정 방법을 기술한다. 다시 말해, 하위개념의 CW값에 상위개념의 CW값을 더하여 하위개념의 CW값을 재조정한다. 과정 2의 처리는 다음과 같이 구성된다.

(1) U-WIN의 계층구조를 이용하여 개념들 사이의 상/하위 관계를 파악한다.

$$if(C_i\, subconcept of\, C_k)$$

(2) 아래와 같이 (1)의 조건을 만족하는 개념들의 상위개념 CW값을 하위개념 CW에 누적하고 집합 CS_i를 구한다.

$$CW(C_i) = CW(C_i) + CW(C_k)$$

$$CS_j = \{C_1, C_2, C_3, \ldots, C_n\} \cap \{C_i, C_k\}^C$$

(3) (2)의 과정을 통해 재설정된 집합을 이용하여 식 (4)를 통해 만족하는 개념을 주제로 선정한다.

$$if(CW(C_m) \geq \sum CW(CS_j)) \quad (4)$$
$$then(Topic = C_m)$$

그림 3은 본 과정에서 문서의 주제를 선정하기 위한 방법을 설명하고 있다.

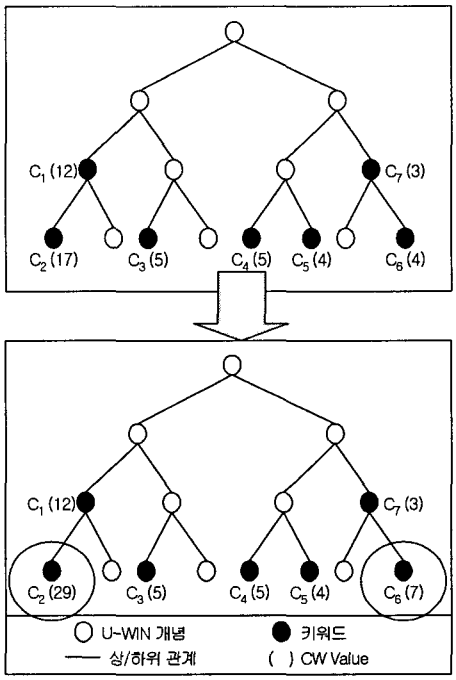

그림 3. 과정 2를 이용한 주제 선정 방법
Fig. 3. Topic Selection Method using Process 2

그림 3의 상단 계층구조와 같은 문서가 있을 때 과정 1을 만족하지 못한다. 이는 과정 2를 거쳐, 상위개념에서 하위개념으로 상속하여 문서의 주제를 선정할 수 있다. 과정 2를 통해 그림 3의 하단 계층구조와 같이 각 개념의 CW값을 재할당하고 식 (4)를 적용하여 주제를 선정하는 것을 표 5에서 보이고 있다.

표 5. 과정 2를 이용한 주제 선정
Table. 5 Topic Selection using Process 2

개념	$CW(C_m)$	$\sum CV(CS_i)$	결과
C_2	29	21	주제선정
C_6	7	43	X

표 5에서 개념 C_2가 식 (4)를 만족하여 주제로 선정됨을 보이고 있다. 본 과정을 통해 만족하는 개념이 존재하지 않을 경우 계속해서 문서의 주제를 선정하기 위해 과정 3으로 진행을 한다.

3.2.3 주제 선정 : 과정 3

본 논문에서는 문서분류를 위한 주제선정을 위해 3가지 경우를 고려하는데, 과정 1과 2에서 선정이 되지 않았을 때는 마지막으로 과정 3에서 결정이 된다. 과정 3은 여러 개의 하위개념을 이용하여 상위개념을 기술하는 상황에 적합한 것으로, 하위개념들의 교집합은 상위개념이 된다는 논리를 전제로 하고 있다. 즉, 일반적인 개념을 설명하기 위해 구체적인 개념들을 이용한 문서에 적합하다. 본 과정에서는 과정 2와는 반대로, 하위개념의 CW값을 상위개념의 CW값으로 역상속을 통해 상위개념의 CW값을 재조정한다. 다음은 과정 3에 대한 구체적인 절차이다.

(1) 추출된 키워드 그룹에서 개념 C_k이 다수의 하위개념 $\{C_j\}$을 갖는다면, C_k의 CW값은 모든 하위개념 $\{C_j\}$의 CW값을 포함한다.

$$if(\{C_j\}\, is\, kind\, of\, C_k)$$

$$then\, CW(C_k) = CW(C_k) + \sum CW(C_j)$$

(2) (1)을 통해 개념들의 CW값을 재할당하고, 집합 CS_i를 새롭게 구한다.

$$CS_i = \{C_1, C_2, C_3, \ldots, C_n\}$$
$$CS_i = CS_i - \{C_j\}$$

(3) 추출된 키워드 그룹의 계층에서 식 (5)를 만족하는 최하위계층을 찾을 때까지 상위계층으로 올라가며 반복한다. 식 (5)에서 CS_i은 (2)의 과정을 통해 얻은 CS_i에서 특정 개념 C_k를 제외한 집합이다.

$$if(CW(C_k) \geq \sum CW(CS_i)) \quad (5)$$
$$then\,(Topic = C_k)$$

과정 3에서, 추출된 키워드 그룹에 없는 개념일지라도 반드시 하나의 개념을 주제로 선정하게 된다. 그림 4와 표 6은 과정 3을 이용해 주제를 선정하는 방법을 보이고 있다.

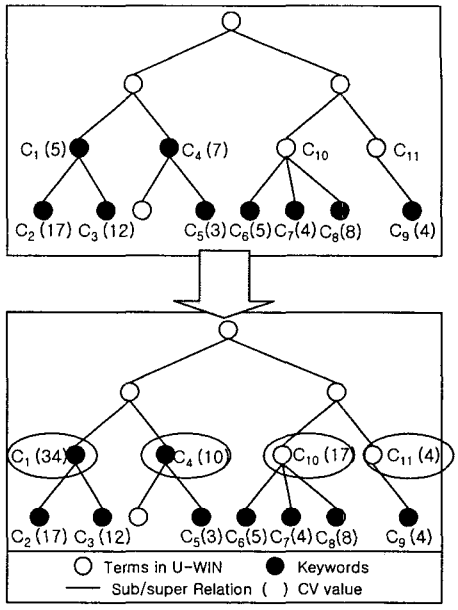

그림 4. 과정 3을 이용한 주제 선정 방법
Fig. 4. Topic Selection Method using Process 3

표 6. 과정 3을 이용한 주제 선정
Table. 6 Topic Selection using Process 3

개념	$CW(C_k)$	$\sum CV(CS_i)$	결과
C_1	34	31	주제선정
C_4	10	55	X
C_{10}	17	48	X

과정 1, 2, 3을 통해 반드시 문서의 분류를 위한 주제 개념이 선택이 되며, 이를 증명하기 위해 실험을 통하여 실제 문서의 분류에 적용하였다.

Ⅳ. 실험 및 평가

본 논문의 문서분류를 위한 주제선정방법의 효율을 보이기 위해, 웹에 있는 국내문서들을 위주로 실험을 하였다. 구글(google.com) 검색 엔진에서, 특정 질의어를 이용하여 검색된 문서를 실험에 사용하였으며, 표 7의 내용은 그중의 하나이다.

표 7. 샘플 문서
Table. 7 Sample Document

싱싱하고 푸짐한 조개를 직접 구워먹는 모듬조개구이 둘이 먹다 죽어도 모를 조개좌판 쟁반가득 푸짐한 모듬조개구이 제부모세마을에 들어서면 죽 늘어선 조개좌판의 진풍경을 볼 수 있다. 제부모세마을 인근 해역에서는 바지락, 맛조개, 죽, 동죽, 소라, 키조개, 홍합, 대합, 피조개 등 다양한 조개들이 풍부하게 서식하고 있다. 막 잡아올린 조개를 큼직한 바구니에 내어오는 모듬조개구이는 보기에도 먹음직스럽다. 연인이나 친구, 가족단위로 서해안의 독특한 정취를 맘껏 느끼며, 싱싱한 조개를 직접 구워 정성들여 만든 소스를 발라 굽거나 혹은 구워서 소스를 찍어 먹으면 행복이 두 배로 늘어난다.

...

집에서 조개구이를 먹을 때는 양념장 재료: 진간장 2큰술, 설탕 2큰술, 고추가루 2작은술, 다진파 2작은술, 깨소금 1작은술, 참기름 1작은술, 다진마늘 1작은술, 실고추, 미나리잎 조금씩 옹기종기 모여 앉아... 조개는 살아 있는 것으로 구입하여 하루밤 정도 소금물에 담가두어 해감시킨 다음 솔로 박박 문질러 씻는다.

...

먼저 핵심 키워드를 추출하기 위해, 표 7의 문서에서 명사개념들을 추출하고, 각 개념의 TF 값과 RV 값을 측정하여 개념들의 그룹들을 식 (2)에 의해 핵심 키워드 그룹을 선정하였다. 표 8은 개념들의 TF, RV값을 통해 식 (2)를 적용하여 키워드 그룹을 선정하는 과정을 보이고 있다.

표 8. 표 7문서의 키워드 그룹 선정 과정
Table. 8 Process of Keyword Group Selection using Table. 7

Group	키워드(TF, RV)	$Max(cwG_j)$
1	조개(7,5),홍합(1,1),키조개(1,1),맛조개(1,1),백합(1,3),소라(1,1),바지락(1,1),대합(1,1)	44
2	그릇(1,2),쟁반(1,1),바구니(1,1)	4
3	간장(1,2),진간장(1,1),양념장(1,1)	4
4	음식(1,1),구이(6,1)	7
5	물(1,2),국물(2,1),소금물(1,1)	5

표 7의 샘플 문서에서 그룹 1을 식 (2)에 의해 핵심 키워드 그룹으로 선정하였다. 선정된 키워드 그룹은 그림 5와 같이 U-WIN의 개념들의 계층구조에 매핑이 되고, 문서분류를 위한 주제선정 과정 1,2,3을 거치게 되는데, 표 9와 같이 과정 1의 조건을 만족하여 '조개'로 문서의 주제를 선정함을 볼 수 있다.

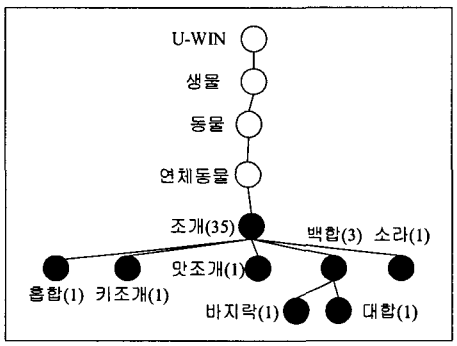

그림 5. U-WIN의 계층구조에 매핑
Fig. 5. Mapping into Hierarchy of U-WIN

표 9. 그림 5에 대한 주제 선정
Table. 9 Topic Selection about Fig. 5

과정	$CW(C_i)$	$\sum(CW(CS_i))$	결과
1	35	9	주제선정

또한 본 연구의 유효성을 평가하기 위해, 웹에 있는 1000개의 국내문서를 임의로 선정하여 문서의 주제선정을 통해 문서분류의 정확성을 테스트하고 통계를 내었다. 문서의 분류를 위해 U-WIN의 계층구조에서 계층 3에 있는 개념들을 이용하여 분류범주(카테고리)를 정하였다.

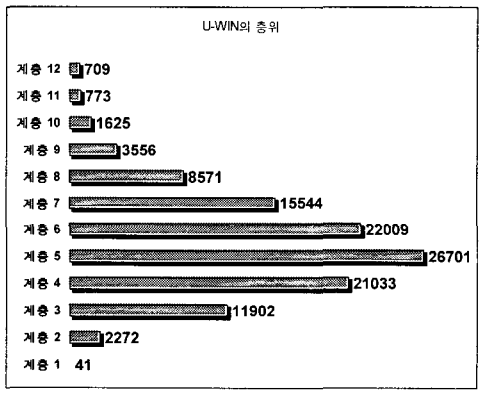

그림 6. U-WIN의 층위별 노드 수
Fig. 6 Nodes Count of Hierarchy Depth

U-WIN의 개념구성은 'UWIN'을 최상위 개념으로 두고, 그림 6과 같이 1계층 41개, 2계층 2272개, 3계층 11902개 개념 등 총 12계층으로 구성되어 있다[11]. 이들 중 계층 3을 일괄적으로 문서분류를 위한 범주로 선정하였다. 표 10은 계층 3에 포함된 개념의 일부를 보이고 있다.

표 10. U-WIN의 계층 3의 개념
Table. 10 Concepts of Third Depth of Hierarchy

포유류, 사람, 공간, 땅, 거주지, 길거리, 편문, 공상, 사랑, 작품, 기구, 가속기, 가압처리장치, 감산기, 감속장치, 개찰기, 거리측정장치, 검출기, 게이트웨이, 경보기, 계산기, 기계, 기화기, 난방장치, 냉각장치, 냉장고, 노즐, 도르래, 변압기, 브레이크, 비디오, 선풍기, 송신장치, 수도꼭지, 수신기, 신호기, 안전장치, 예열기, 온수기, 온장고, 와이퍼, 잭, 전자장치, 제어장치, 증류장치, 축전기, 충전기, 카세트, 콘솔, 튜너, 프로펠러, 하드웨어, 핵탄두 ...

표 10에 있는 개념들을 이용해 분류범주를 두고 본 연구에서 제안하는 방법을 실험한 결과 각 영역에 정확하게 분류한 정확도가 84.4%를 보였다. 이는 Sinka[14]의 연구에서 제공되는 도메인 온톨로지를 기반으로 연구된 온톨로지 기반의 분류 방법[13]과 베이지언 분류법 (naive Bayesian classifier)[15]를 이용하여 비교해볼 때, [13]은 0.4%, [15]는 11.4% 향상된 정확도를 보이지만, 카테고리의 수를 볼 때 더욱 세분화된 분류가 가능하다 할 수 있다. [13]과 [15] 방법의 정확도는 표 11과 같다.

표 11. 비교평가
Table. 11 Evaluation

방법 (카테고리 수)	정확도 (%)
온톨로지 기반 분류방법 (14)	84.0%
베이지언 방법 (14)	73.0%
제안된 방법 (11902)	84.4%

또한 U-WIN에 분류체계를 분석해 볼 때, 계층 3 또는 계층 4까지만 이루어져 있는 개념도 다수 존재하였으며 이들의 분류는 계층 2에 있는 개념들로 10.3% 분류가 되었다. 이는 분류범주에 포함되지는 않지만 의미적, 개념적 측면으로 판단할 때 본 연구가 합당하다는 결론을 맺을 수 있다.

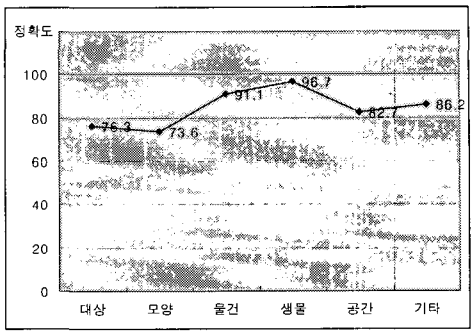

그림 7. 계층영역별 실험결과
Fig. 7. Experiment Result classified by Domain

또한, U-WIN에는 'UWIN'을 최상위로 두고 41개의 영역을 갖고 있는데, '물건', '생물'의 계층영역은 계층구조가 명확하고 정의된 개념들이 일반 문서에도 자주 사용되어 다른 도메인에 비해 주제선정의 정확도가 높은 것으로 나타났다. 또한 '대상', '모양', '공간'등의 도메인은 계층을 구성하는 개념들이 자주 사용되지 않는 개념들이 많았으며, 일부 개념들은 U-WIN에 정의만 되어있고 계층구조를 이루지 않는 개념들이 존재하여 그림 7과 같이 도메인에 따라서 정확도의 차이가 컸다.

전체 정확도 84.4%는 각 계층영역별로 '대상' 76.3%, '모양' 73.6%, '물건' 91.1%, '생물' 96.7%, '공간' 82.7로 구성되었으며, 상기 도메인을 제외한 나머지에서 86.2%를 구성하였다. 또한 실험결과 키워드 추출, 주제선정의 잣대가 되는 U-WIN의 개념들 관계가 상/하위관계, 유의어관계만 작성되어, 문서내의 부분/전체(표 8에서 그룹 1과 4는 관계가 존재할 수 있음)가 되는 개념들 사이의 관계를 파악하지 못하여 본 논문의 핵심 키워드를 추출이 부족하다는 한계가 존재하였다.

V. 결론 및 향후 연구

본 논문의 핵심은 웹에 산재되어 있는 한글문서들을 자동으로 분류하기 위해, 문서의 주제를 선정하기 위한 방법을 제안하고 있다. 문서의 주제선정은 문서내에 포함된 개념들의 TF(Term Frequency)와 개념들 사이의 관계 횟수(RV : Relation Value) 그리고 한글의 개념과 관계를 파악하여 표현한 형식적이고 명세적인 어휘 네트워크인 U-WIN을 이용하고 있다. 또한 이들을 기반으로

핵심 키워드가 되는 그룹을 추출하는 방법, 추출된 키워드들을 바탕으로 3가지 과정을 거쳐 주제를 선정하는 방법을 제안하고 있다. 실험결과에서 정확도는 84.8%로 효율적임을 증명하였으며, 국내 어휘 사전인 U-WIN을 이용하여 한국어 문서분류를 위한 시도라는 점에 의의가 있다.

국내문서의 주제선정에 있어서 개념들의 관계를 문서 자체 내에서 찾아내는 방법, 더욱 정확한 핵심 키워드 추출 등은 문서분류의 정확도를 높이기 위해 꾸준히 진행되어야 할 과제이다.

참고문헌

[1] Jinze Liu, Wei Wang, Jiong Yang, "Research track posters: A framework for ontology-driven subspace clustering", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining KDD '04, pp. 623-628, ISBN:1-58113-888-1, Aug. 2004.

[2] Illhoi Yoo, Xiaohua Hu, "A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE", International Conference on Digital Libraries archive Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries table of contents, pp. 220-229, ISBB:1-59593-354-9, 2006.

[3] Hwanjo Yu, ChengXiang Zhai, Jiawei Han, "Text classification from positive and unlabeled documents", Source Conference on Information and Knowledge Management archive Proceedings of the twelfth international conference on Information and knowledge management, ISBN:1-58113-723-0, pp.232-239, 2003.

[4] Thierson Couto, Marco Cristo, Marcos André Gonçalves, Pável Calado, Nivio Ziviani, Edleno Moura, Berthier Ribeiro-Neto, Belo Horizonte, "A comparative study of citations and links in document classification", Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, ISBN:1-59593-354-9, pp.75-84, 2006

[5] Yifen Huang, Tom M. Mitchell,"Text clustering with extended user feedback", Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 413-420, ISBN:1-59593- 369-7, 2006.

[6] Hyunjang Kong, Myunggwon Hwang, Gwangsu Hwang, Jaehong Shim, Pankoo Kim, "Topic Selection of Web Documents Using Specific Domain Ontology", MICAI 2006: Advances in Artificial Intelligence, LNAI 4293, pp.1047-1056, 2006.

[7] Greiner, R., Grove, A, Schuurmans, D.: On learning hierarchical Classifications (1997)

[8] Quek, C.Y, Mitchell, T: Classification of World Wide Web Documents. Seniors Honors Thesis, School of Computer Science, Carnegie Melon University (1998)

[9] Koller, D., Sahami, M.: Hierarchically Classifying Documents Using Very Few Words. In the Proceeding of Machine Learning (ICML-97) (1997) 170-176

[10] http://en.wikipedia.org/wiki/Tf-idf

[11] 김준수, 옥철영, "정제된 의미정보와 시소러스를 이용한 동형이의어 분별시스템", 정보처리학회논문지 B 제12-B권 제7호 pp.829~840 2005. 12.

[12] 허준희, 최준혁, 이정현, 김중배, 임기옥, "문서의 주제어별 가중치 부여와 단어 군집을 이용한 한국어 문서 자동 분류 시스템", 정보처리학회논문지 B 제8-brnjs 제5호 pp.447-454 2001.10.

[13] 최재혁, 서혜성, 노상욱, 최경희, 정기현, "온톨로지 기반의 웹 페이지 분류 시스템", 정보처리학회논문지 B 제 11-Brnjs, 제 6호, pp723-734, 2004년 10월.

[14] M.P.Sinka and D.W.Corne, "A large benchma가 dataset for web document clustering," Soft Computing Systems:Design, Management and Applications, Frontiers in Artificial Intelligence and Applications, Vol.87, pp.881-890, 2002.

[15] R.Hanson, J.Stutz and P.Cheeseman, "Bayesian Classification Theory", Techinical Report FIA-90-12-7-01, NASA Ames research Center, AI Branch, 1991.

[16] 황명권, 배용근, 김판구, "문서 내용의 계층화를 이용한 문서 비교 방법", 한국해양정보통신학회논문제 제10권 12호, pp2335-2342, 2006년 12월

저자소개

고 광 섭(Kwang-Sup, Ko)

1989 영국 워릭대학(경제학석사)
2003. 3-현재 건국대학교
컴퓨터공학과 박사과정

※관심분야 : 문서분류, 전자정부, 정보통신정책

황 명 권(Myung-Gwon, Hwang)

2004 조선대학교 컴퓨터공학부
(공학사)
2006 조선대학교 대학원 전자계산학과
(이학석사)
2006.3-현재 : 조선대학교 대학원 컴퓨터공학부 박사과정
※관심분야 : 문서분류, 시맨틱웹, 멀티미디어검색

김 판 구(Pan-Koo, Kim)

1988 조선대학교 컴퓨터공학과
(공학사)
1990 서울대학교 대학원 컴퓨터공학과
(공학석사)
1994 서울대학교 대학원 컴퓨터공학과 (공학박사)
1995-현재 조선대학교 컴퓨터공학부 교수
※관심분야 : 시맨틱웹, 온톨로지, 멀티미디어 정보검색,
감성정보처리

이 창 훈(Chang-Hoon Lee)

1977 한국과학기술원 전산학과 (석사)
1993 한국과학기술원 전산학과 (박사)
1980-현재 건국대학교 컴퓨터공학과
교수
※관심분야 : 지능시스템, 보안, 전자상거래 등