효율적인 자원 운영을 위한 전문용어 생명주기 관리 연구

정한민°* 구희관** 이병희* 성원경*
*한국과학기술정보연구원 차세대정보시스템연구실
**과학기술연합대학원대학교 응용정보과학전공
{jhm°, hkkoo, bhlee, wksung}@kisti.re.kr

Study of Term Life Cycle Management for Efficient Resource Use

Hanmin Jung°*, Hee-Kwan Koo**, Byeng-Hee Lee*, and Won-Kyung Sung*
*Information System Research Lab., Korea Institute of Science and Technology Information
**Practical Information Science, UST

요 약

전문용어는 사전, 시소러스 및 온톨로지를 비롯한 다양한 기반지식자원에서 사용되고 있으며, 해당분야 발전에 민감하게 반응하는 특징을 가진다. 그럼에도 불구하고, 용어를 생명주기 관점에서 바라보고 이를 효율적으로 사용할 수 있도록 하는 연구가 부족하다. 본 논문에서는 한정된 인적·물적 자원을 효율적으로 사용할 수 있도록 가까운 미래에도 유용한 용어들을 선정하고 관리하기 위한 방안으로서 용어지배값(TDV; Term Dominance Value)을 제안한다. 이를 통해 용어 생명주기의 각 단계를 생성, 성장, 유지, 쇠퇴, 소멸, 재생 등으로 정의함으로써 관리해야 할 대상 용어를 명확히 할 수 있도록 한다. 용어지배값과 Coverage와의 관계 실험을 통해 유망한 용어들을 선정하고 관리해야 하는 당위성을 보여준다.

1. 서 론

전문용어는 사회적 언어현상뿐만 아니라 기술발전과도 밀접한 관계를 가진다. 특히, IT 분야의 경우에는 용어들의 생명주기가 짧으며, 생명주기 상에서 유형변화가 심하게 일어난다. 예를 들어, 전자신문 말뭉치를 대상으로 살펴보면 2002년 용어빈도 기준 상위 30개 복합어들 중 2003년도에는 19개, 2003년 상위 30개 중 2004년도에는 21개만이 그 순위를 유지함을 알 수 있다. 또한, 2004년 말뭉치에서 자동 추출한 59,022개 복합어들 중 25,957개는 1998 ~ 2002년도에 출현하지 않은 신조어에 해당하는 용어후보들이다. 그렇지만, 현재 전문용어사전들(KORTERM, TTA 등)은 그 구축과정에서 용어 선정시의 우선화 원칙과 Coverage에 대한 충분한 고려가 없었으며, 실제 이들을 포함한 약 20만 사전 엔트리를 2004년도 전자신문 말뭉치에 적용하였을 때 약 7.5%의 Coverage만을 보여주고 있다.

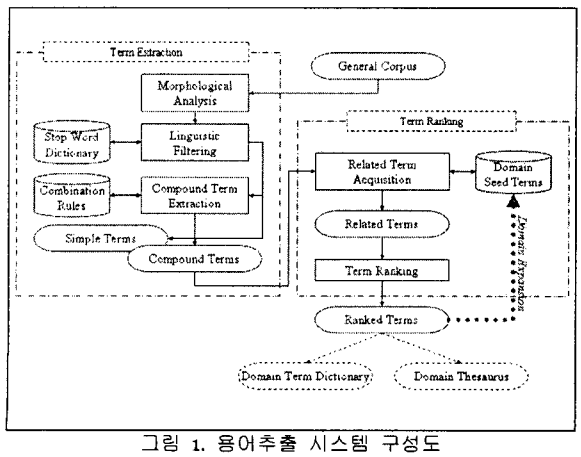


그림 1. 용어추출 시스템 구성도

본 연구에서는 이러한 특성을 갖는 용어들을 용어후보 추출, 용어후보 순위화, 용어 선정 및 용어 생명주기 관리를 통해 일관성 있게 다룸으로써 한정된 인적·물적 자원을 효율적으로 사용할 수 있는 방안을 제시하고자 한다. 이러한 연구는 현재뿐만 아니라 가까운 미래에도 그 효율성을 보장할 수 있도록 함으로써 실효성 있는 기반지식자원 확충이 가능하도록 한다.

2. 용 어 추 출 및 용 어 선 정

본 논문에서 사용한 용어들은 자동으로 이루어지는 용어후보추출과 수동으로 이루어지는 용어선정을 통해 얻어진다. 용어후보추출은 형태소 분석, 불용어 필터링, 복합어를 위한 결합규칙 처리 및 용어후보 순위화 과정을 포함한다.

전문용어에 대한 정의와 기준을 제시한 연구들이 있었음에도 불구하고 해석의 어려움과 원칙론적 제시로 인해 용어 선정을 위한 명확한 가이드라인이 되기에는 부족하다 [3]. 본 연구에서는 자동 추출한 용어후보들에 대해 재정의된 다음의 기준들을 이용하여 작업자들이 보다 명확히 전문용어를 선정할 수 있도록 하고자 하였다 (본 연구의 후속 연구가 범용 과학기술분야에 대한 계층적 개념망/어휘망 구축이므로 이에 부합하는 전문용어를 선정하는 것을 목표로 하고 있다.). 본 실험에서 사용한 용어들도 이러한 선정과정을 거치고 2인 작업자가 상호검사하여 선택된 용어들이다.

a. *영어와 혼용된 형태도 전문용어로 인정한다 (예. "e러닝," "DVD플레이어").*
b. *음차표기 용어도 전문용어로 인정한다 (예. "더블데이터레이트," "유비쿼터스네트워크").*
c. *PLO를 제외한 고유명사도 전문용어로 인정한다 (예. "SQL서버," "MSN메신저").*
d. *사전에 등록된 용어라도 쓰이는 분야가 명확한 경우는 전문용어로 인정한다 ("프린터," "모니터").*
e. *분야 파악이 용이한 용어도 전문용어로 인정한다*

(예. "생체인식," "공인인증").

f. 일반적 서술형 명사를 포함하는 용어는 전문용어에서 제외한다 (예. "시스템구현," "S/W개발").

g. 조어생성에 제약이 약한 용어는 전문용어에서 제외한다 (예. "음악/문서/멀티미디어파일," "위치/서지/지식정보").

3. 용 어 지 배 값

우리는 용어 생명주기 관리를 위해 용어지배값 (TDV; Term Dominance Value)을 정의하여 사용한다. 용어지배값은 성장하고 있거나, 쇠퇴하고 있는 정도를 일정기간 내에서의 관찰을 통해 수치화한 값으로 이 값의 추이에 따라 용어의 현재 단계를 추정할 수 있다. 시작 (start)과 종료 (end) 시점 사이의 기간은 TDV 추이를 일관성 있게 관찰하기 위해 고정하며, 본 논문에서는 5년을 사용하고 있다.

$$TDV(t)_{start\text{-}end} = \frac{\sum_{i=start}^{end}((NTF_i - ANTF_t) * PW_i^2)}{PF_t}$$

$TDV(t)_{start\text{-}end}$: 용어 t에 대해 start부터 end까지 관찰하여 얻은 용어지배값

NTF_i: 특정 시점 i에서 발생한 정규화된 용어빈도 (term frequency of time i * (minimum corpus size / corpus size of time i))

$ANTF_t$: 용어 t에 대해 start부터 end까지 발생한 정규화된 용어빈도의 평균

PW_i: 특정 시점 i에 대한 기간 가중치 (0.1 * (time i – start) + 0.1)

PF_t: 용어 t에 대해 출현한 시점빈도

4. 용어 생명주기

[1]은 지식 생명주기를 Creation, Mobility, Diffusion 및 Commoditization으로 구분하고 있으나, 사용하지 않는 지식을 버리고 새로운 것을 추가하는 등의 자원관리 측면에서 본다면 Termination을 추가할 필요가 있다 [2]. 전문용어 관점에서 본다면 그 생명주기가 지식보다 상대적으로 짧을 수 밖에 없으므로, 우리는 한정된 자원을 효율적으로 사용하기 위해 쓰임새의 변화가 심한 용어 특성을 반영하여 용어 생명주기를 재정의하고 이를 이용한 관리방안을 본 장에서 제안한다.

한 용어에 대해 일정기간 내에서 관찰한 두 TDV의 변화형태에 따라 해당 용어의 단계를 다음과 같이 나눌 수 있다. ε는 Error Rate이며, 본 논문에서는 0.0과 0.1을 사용하고 있다.

용어의 생명주기는 표 1과 같이 생성, 성장, 유지, 쇠퇴, 소멸, 재생의 단계로 정의할 수 있다. 물론, 유지기간이 아주 짧을 수도 있으며, 성장 없이 바로 쇠퇴할 수도 있다. 또한, 이형태를 많이 가지는 용어들 (예. "NAND플래시," "낸드플래시," "난드플래시," "낸드형플래시" 등)은 서로간의 생존 경쟁 과정 속에서 급격한 생명주기를 가지게 되는 경우가 많다.

표 1. TDV 변화형태에 따른 생명주기 상에서의 단계

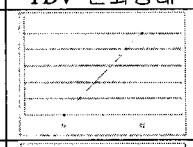
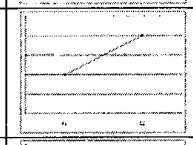
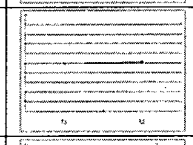
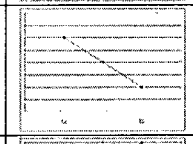
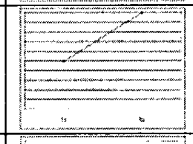
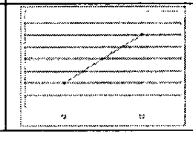
생명주기 단계	TDV 변화형태	비고
생성 (新造語)		$TDV_{t1}=0$ and $ACCTF_{t1}=0$[1]
성장 (Growing)		$\lvert TDV_{t2} - TDV_{t1}\rvert > \varepsilon$
유지 (Steady)		$\lvert TDV_{t2} - TDV_{t1}\rvert \leq \varepsilon$
쇠퇴 (Declining)		$\lvert TDV_{t2} - TDV_{t1}\rvert > \varepsilon$
소멸 (死語)		$TDV_{t2}=0$ and $ACCTF_{t2}=0$
재생 (Recycling)		$TDV_{t1}<0$ and $TDV_{t2}>0$

그림 2는 일반적인 용어 생명주기에 따른 TDV 추이를 보여준다. 생성 이전에는 0의 값을 가지다가, 그 값이 성장 단계에서는 점점 커지게 된다. 쇠퇴과정에서는 다시 하락하게 되며, 음수 영역에서 소멸되는 순간 0으로 값으로 되돌아간다. 재생형 용어는 용어가 가지는 개념이 변이되거나, 타 분야로 전이되거나, 다시 이슈로 등장하는 경우에 나타날 수 있다. 이에 대한 연구는 향후 연구에서 다루고자 한다.

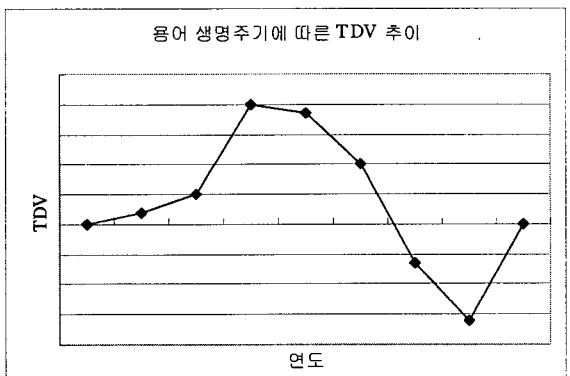


그림 2. 용어 생명주기에 따른 TDV 추이

[1] $ACCTF_{t1}$: TDV_{t1}이 관찰한 기간 내에서의 누적 용어빈도 수

용어 생명주기 관찰을 위해 직전 연도의 용어빈도나 누적빈도 등을 자질로 사용할 수도 있다. 그러나, 직전 연도의 용어빈도는 그 이전기간에서 용어가 어떻게 출현하였는 지를 알 수 없으며 신조어와 사어에 대한 판단을 하기 어렵다. 누적빈도는 단지 현재까지의 전체 출현빈도만을 고려함으로써 용어의 출현빈도 추이를 관찰할 수 없다는 문제점을 가진다.

본 연구에서 제안하는 TDV는 용어 생명주기를 직관적으로 보여줌으로써 현재 용어의 단계를 용이하게 판단할 수 있도록 해준다. 또한, 신조어와 사어를 구분하고 이를 자원의 효율적인 활용에 이용함으로써 가까운 미래에 지배도 높은 용어들을 우선적으로 고려할 수 있도록 해준다.

5. 실 험

그림 3과 4는 상·하위 10위에 속하는 성장형 용어들과 쇠퇴형 용어들의 연도별 TDV 추이를 보여준다. 성장형 용어들 중 "위성 DMB"와 "지상파 DMB"를 포함하는 5개 용어들은 2003년에 처음 출현한 신조어들이다.

2002년에 67,259번, 2003년에 87,676번, 2004년에 249,360번의 전체 복합어 용어빈도를 가지며, 이들 용어들에 대한 전체 용어빈도 대비 Coverage를 계산한 결과가 표 2이다. 성장형 용어의 경우 최근 연도일수록 Coverage가 상승하는 반면, 쇠퇴형 용어의 경우 2002년에 성장형 용어보다 Coverage가 높았음에도 불구하고 최근 연도일수록 Coverage가 하락한다. 이를 통해, 성장형 용어의 중요성이 가까운 미래에도 유지될 수 있으므로 한정된 자원 내에서는 이들에 대한 고려를 높여야 함을 알 수 있다.

표 2. 전체 용어빈도 대비 성장형 용어들과 쇠퇴형 용어들의 Coverage 추이
($TDV_{2000-2004}$ 기준 상·하위 10개 용어, 복합어)

연도	성장형 용어 (%)	쇠퇴형 용어 (%)
2002	0.2497807	2.2227509
2003	1.4268443	1.1713582
2004	2.4366378	0.8930863

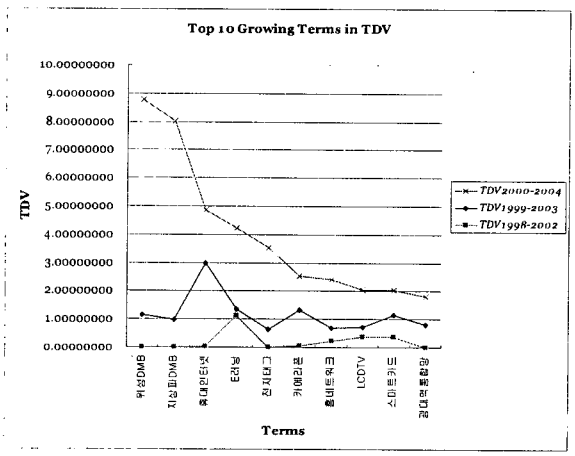


그림3. 상위 10위에 속하는 성장형 용어들에 대한 TDV 추이 ($TDV_{2000-2004}$ 기준, 복합어)

TDV가 상위에 속할수록 성장형 용어들이 많으며, 하위에 속할수록 쇠퇴형 용어들이 많다는 사실도 이번 실험을 통해 확인할 수 있었다. ε=0.1인 환경에서는 $TDV_{2000-2004}$ 기준 상위 30개 용어들 중 28개가, ε=0.0인 환경에서는 26개가 성장형 용어들이다. 반면에, ε=0.0과 ε=0.1인 환경에서 $TDV_{2000-2004}$ 기준 하위 30개 용어들 중 26개가 쇠퇴형 용어들이다 (상위 30개 용어들 중 상승형이 아닌 나머지 4개는 진동형이며, 하위 30개 용어들 중 쇠퇴형이 아닌 나머지 3개는 진동형이며, 1개는 상승형이다.).

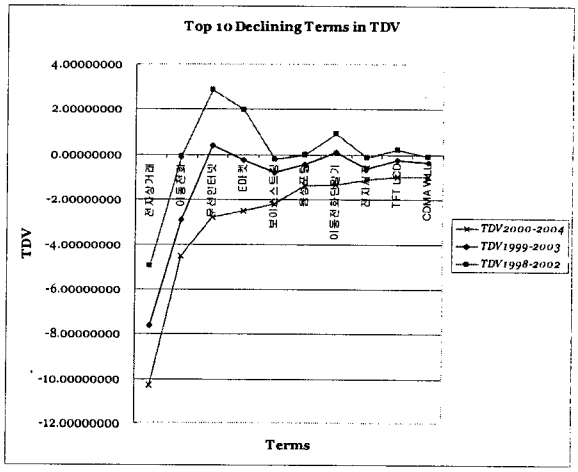


그림4. 하위 10위에 속하는 쇠퇴형 용어들에 대한 TDV 추이 ($TDV_{2000-2004}$ 기준, 복합어)

6. 결 론

본 논문은 한정된 자원을 효율적으로 이용하기 위해 용어지배값 (TDV; Term Dominance Value)과 용어 생명주기를 정의하고, 용어 생명주기에 따른 관리방안을 제시하고 있다. 본 연구를 통해 제안하는 TDV는 용어 생명주기 변화를 관찰하기에 좋은 하나의 기준임을 상기 실험이 보여주고 있다. 그렇지만, 현실적으로 TDV에만 의존하게 되면 쇠퇴형 용어들 중에서 Coverage가 큰 것들에 대한 고려가 약해질 수밖에 없다. 가까운 미래에 쇠퇴형 용어들의 영향력이 점진적으로 약해지겠지만 Coverage를 고려하지 않을 수 없으므로 TDV 이전에 기본적인 Coverage 비율을 정하고 이들 내에 포함되는 고빈도 용어들은 우선적으로 구축할 필요가 있다. 향후 연구에서는 최적화된 자원 이용이 가능하도록 Coverage와 TDV 사이의 관계를 명확히 규명할 예정이다. 또한, 재생형 용어에 대한 사회적 현상, 분야 전이 등도 연구 대상이다.

참 고 문 헌

[1] J. Birkinshaw and T. Sheehan, *Managing the Knowledge Life Cycle*, Journal of MIT Sloan Management Review, Vol. 44, No. 1, 2002.

[2] P. Feher, Knowledge Termination: The End of the Game, *Proceedings of Information Technology Interfaces*, 2003.

[3] 최기선, 송영빈 외, *전문용어연구1: 한국에 있어서의 전문용어 연구와 방향 (2편 전문용어학의 제문제)*, 홍릉과학출판사, 2000.