분야 연상어를 이용한 패시지검색 방법

장기철, 정규철, 이윤정, 박기홍
군산대학교 컴퓨터정보과학과
e-mail : kcjang@kunsan.ac.kr

Passage Retrieval Using Field Reminding Words

Ki-Cheol Jang, Kyu-Cheol Jung, Yoon-Jung Lee,
Kihong Park
Dept of Computer Information Science, Kunsan Natl University

요약

인터넷의 보급에 의해, 전자화된 문서가 대량으로 유통되게 되었다. 이에 따라, 대량의 전자화 문서로부터 검색 요구에 해당하는 문서를 검색한 기술이 요구되고 있다. 문서 검색시 복수의 화제나, 분야가 혼재한 문서로는, 검색 요구에 해당하는 내용이 문서의 일부분에 존재하는 경우가 대부분이다. 이처럼 문서전체를 검색 대상으로 하는 것이 아니라 검색 요구에 일치한 일부만을 검색한 패시지 기술이 주목되고 있다. 본 논문에서는 패시지가 있는 분야의 화제에 관하여 쓰여진 정리로서 파악하고 분야 연상어를 이용하고, 패시지를 결정한 방법을 제안하였다. 결정된 패시지와 미리 사람에 의하고 결정한 패시지가 어느 정도 일치하고 있는지를 비교하여 제안 방법의 유효성을 평가한 결과 적합율이 약 80%, 평균 재현율이 약 67%되어, 유효성을 확인할 수 있었다.

1. 서론

최근 인터넷의 보급에 의해, 전자화된 문서가 대량으로 유통되게 되었다. 이에 따라, 대량의 전자화 문서로부터 검색 요구에 해당하는 문서를 검색한 기술이 요구되고 있다.

본 논문에서 다루게될 패시지 검색이란 문서 전체를 하나의 단위라고 간주하는 검색과는 다르게, 패시지라는 단위를 사용하고, 검색 요구와 문서와의 유사도 계산을 한다. 패시지란 일반적으로는 문서 중에 연속적인 일부분에 관한 것을 말한다. 그러나, 패시지 검색에 있어서는 단지 연속한 일부분이라고 말한 것만으로는 불충분하고, 문서 중에 검색 요구가 강하게 관련된 의미적인 정리를 형성할 필요가 있다.

이전의 패시지 검색으로서는, 검색해야 할 패시지의 단위에 의해, 문서내의 장·절이나 형식 단락과 같은 형식적인 정보에 근거한 것[1,2,3,4]과, 고정길이나 가변길이의 윈도우에 근거한 [5,6]한 것과, 형식에 의하지 않는 의미적인 정리에 근거한 것 [7]등이 제안되고 있다.

한편, 분야 연상어란 , "투수" 나 "선거" 같이 단어만 봐도, "야구" 나 "정치" 라는 상식적 분야를 인지한 것이 가능한 단어 또는 복합어에 관한 것을 말한다. [8]들은 고정된 분야 체계에 따라 분류된 모든 자료로부터 각 분야 특유의 분야 연상어를 구축한 수법을 제안하고 있다.
분야 연상어가 나타나는 주변의 부분은, 그것이 나타내는 분야의 화제라고 생각된다. 그러나, 분야 연상어가 나타나지 않는 패시지 (passage) 에 대해 어떻게 분야를 결정하든지가 문제로 대두된다. 그러면, 화제의 흐름의 특징을 검증하고, 분야 연상어의 연속 출현율을 기초로 화제의 계속성을 계산하고, 화제의 전환성을 정의한 것으로, 효율적으로 패시지를 결정한다

그리고, 전문가법에 의해 결정된 패시지와, 미리 사람에 의하고 결정한 패시지가 어느 정도 일치하고 있든 지를 비교하고, 제안 방법의 유효성을 평가했다. 그 결과, 스포츠에 관한 50개 문서의 정답 패시지에 대한 평균 적합율이 약80%, 평균 재현율이 약67%로 되어, 유효성이 확인할 수 있었다.

아래, 제2장에서는 본 논문의 연구 대상으로 되는 패시지 검색의 이전방법에 관하여 설명하고, 문제점에 대해 논한다. 제3장에서는, 본 논문에 사용된 분야 연상어에 관하여 설명한다. 제4장으로는, 대상 문서에 대해 분야 연상어를 이용하여 패시지 결정을 한 방법에 관하여 논하고, 제5장에서는, 전문가법을 이용하고 실험을 하고, 전문가법의 유효성을 확인한다. 제6장으로는, 결론과 향후의 과제에 관하여 논한다.

2. 기존의 패시지 검색

2. 1 형식 단락에 근거하는 패시지 검색

이 타입의 패시지 검색에서는, 표제와 형식 단락을 각각 하나의 패시지로서 취급한다. 또, 벡터 공간 모델에 의해 패시지와 검색 요구와의 국소적 유사도를 계산하여, 국소적 유사도의 큰 문서를 검색 후보로 하고 있다. 패시지마다의 각 term의 중요도는, term의 패시지 내에서 출현 빈도 TF 및 term의 문서 집합 전체에서 출현 빈도의 역 IDF 에 근거해, 식(1)으로 계산된다.

$$w(t) = tf(t) \times \log(N / n(t)) \qquad -(1)$$

단, tf(t)는 패시지 안에 출현한 term t의 출현 빈도, N은 모

든 단락이고, n(t)는 term t의 출현 단락의 수이다.
또, 각 패시지를 term의 가중치 벡터 P로 표현하고, 검색 요구 벡터Q와 패시지 벡터P 상이의 유사도는 식 (2)을 요구한다.

$$sim(Q, P) = \Sigma t \left(tf(qt) / \log(N / n(t)) \right)2 \times w(t) \quad -(2)$$

단, tf(qt)는 검색 요구내의 term qt의 검색 요구내의 빈도

형식 단락이나 장·절 같은 형식적인 정보에 의한 패시지 추출 기법은 저자가 결정한 구조에 따라 패시지를 정하는 기법이다. 이 타입의 패시지는 검색에 앞서 인덱싱 할 때 결정할 수 있으므로 처리가 용이 하는 이점이 있다. 그러나, 실재 문서는 동일한 화제가 복수의 단락이나 장·절 등에 걸치는 경우가 많다.

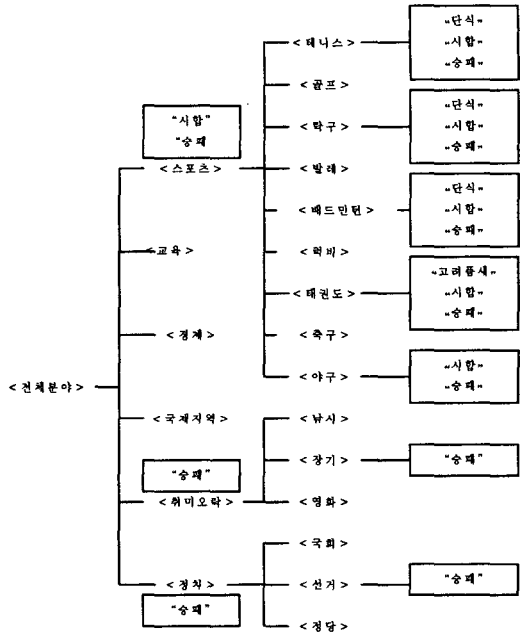
2.2 윈도우에 근거한 패시지 검색
고정길이나 가변길이 윈도우에 의한 패시지는, 검색 요구가 입력된 시점에서 윈도우를 슬라이드 시키면서 각 문제를 주사하여, 검색 요구와 유사도가 높은 윈도우를 결정한다.
윈도우를 슬라이드 시키는 폭에 관해서는, 슬라이드 폭을 작게 하면 섬세한 주사가 가능해지지만, 검색에 관련된 비용이 높아져 비실용적이 되고, 폭을 크게 하면 검색 비용은 경감되지만, 윈도우 경계에 걸치는 부분의 영향을 의해 검색 정밀도가 나빠지는 가능성이 있다.

2.3 의미적인 정리에 근거한 패시지 검색
형식에 의하지 않는 의미적인 정리에 근거한 패시지 추출은, 문서의 내용에 근거하고 있기 위해 가장 바람직한 방법이다. 이 타입의 패시지 추출에는, 미리 문서를 담화부분으로 분할한 것으로 인덱싱 시에 패시지를 결정한 방법과 검색 요구가 입력된 시점에서 검색요구와 관련이 강한 의미적인 정리의 부분을 패시지로 꺼내는 2가지 방법이 있다. 문서를 미리 고정길이블록의 단락과 문서 중에 출현한 단어의 결속성을 계산하고, 결속성을 갖는 단어가 블록 사이에 걸치고 있는 비율이 많은 블록 끼리를 정리하고, 담화부분, 즉 패시지를 형성한다. 그러나 형식적인 정보를 이용한 경우와 같이, 검색 요구에 관계없이 패시지가 결정되기 위해 어떤 검색 요구에도 적절한 패시지를 가정하고 있다고 말한 문제점이 있다.

3. 분야 연상어
분야 연상어라는 것은 분야를 연상할 수 있는 단어 또는 복합어를 말한다. 또한 분야 연상어는 형태소 연상어와 복합 연상어로서 정의한다. 분야 체계에 의한 분야 트리를 정의하고 그 만큼의 노드에 대해 연상어의 수준을 정의한다. 그 다음에 분야 연상어가 시간 경과에 의해 변화하는 것을 주목하여 안전 랭크를 새롭게 정의한다.

3.1 연상어와 분야 체계
분할이 불가능하고 의미를 가진 최소의 단위를 형태소라고 부르며, 두 단어 이상으로 구성되는 말을 복합어라고 부르며, ""안에 기술한다. 또 형태소의 분야 연상어를 형태소 연상어, 복합어의 분야 연상어를 복합 연상어라 한다. 단, 미등록 어는 연상어의 대상에서 제외한다.
또, 각각 하나의 접사와 명사로 구성되는 일반적인 복합어("소득세", "핵연료", "온난화" 등)는 세분화할 경우 분야 정보를 잃게 되므로 형태소 연상어로 취급한다.

분야 체계는 분야 트리로 구성하여 분야의 터미널에 해당하는 분야를 종단 분야, 종단 분야 이외는 중간 분야라고 부른다. 직접적인 상위 분야를 부모 노드, 하위 분야를 자식 노드라고 부른다. 분야의 지정은 분야명의 패스<S>로 기술하지만, 뿌리에 해당하는 <전체 분야>는 생략한다. 특히 모순이 생기지 않는 경우는 패스 지정을 생략해 종단 분야만으로 설명한다. <>안의 분야명과 구별하기 위해 의미나 개념명 등은 []안에 기술한다.



[그림 1 연상트리와 연상어의 예]

그림 1은 분야 트리의 예를 나타낸다. 예를 들면, 분야 패스 <S>=<스포츠\태권도> 는 <스포츠>의 하위의 종단 분야<태권도>를 표현한다.

3.2 분야 연상어에 있어서의 수준과 안정성 순위
3.2.1 분야 연상어의 수준
단어는 유일한 종단 분야나 중간 분야를 정하는 경우, 또는 복수의 종단 분야나 중간 분야를 정하는 경우가 있으므로 연상어 수준을 다음에 정의한다.

[정의 1] 연상어 w의 분류와 수준
(수준 1) 완전 연상어 w
w는 유일한 종단 분야만을 연상한다.
(수준 2) 준 완전 연상어 w
같은 부모들을 가진 종단 분야 중에서 한정된 복수의 종단 분야만을 연상한다.
(수준 3) 중간 연상어 w
w는 완전 연상어, 준 완전 연상어가 아닌 유일한 중간 분야를 연상한다.
(수준 4) 다분야 연상어 w
w는 완전 연상어, 준 완전 연상어, 중간 분야 연상어가 아닌, 복수의 중간 분야나 종단 분야를 연상한다.
(수준 5) 비연상어 w
w는 수준 1-4이외고, 특정 분야를 연상하지 않는다.

연상어	연상분야	수준
고려품새	<스포츠\태권도>	1
단식	<스포츠\테니스>	2
	<스포츠\탁구>	
시합	<스포츠>	3
승패	<스포츠>	4
	<취미·오락\장기>	
	<정치\선거>	
경우		5

[표1 분야 연상어의 예]

[예 1]
표1은 분야 연상어의 예를 나타낸다. 수준 1의 완전 연상어는, "고려품새"와 같은 종단 분야<태권도>를 한번에 정한다. 수준 2의 준 완전 연상어는 "단식", "복식"과 같이 같은 부모인 <스포츠>내의 복수의 종단 분야 <테니스>, <탁구>, <배드민턴>을 가리킨다.
수준 3의 중간 연상어는, "시합"과 같이, 종단 분야는 특별히 정할 수 없지만, 하나의 중간 분야 <스포츠>를 가리킨다. 또, 수준 4의 다분야 연상어 "승패"는 복수의 종단 분야 <취미·오락\장기>, <정치\선거>나 중간 분야 <스포츠>를 가리킨다. 수준 5의 비 연상어는 " 경우", "사용"과 같이 분야를 가리킬 수 없는 단어이다.

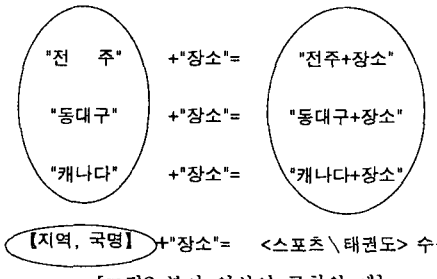
수준	안정성 랭크	연상어 후보	분야
1	b	투원스	<스포츠\야구>
1	a	투수	<스포츠\야구>
1	b	세이브	<스포츠\야구>
1	a	야구	<스포츠\야구>
1	a	홈런	<스포츠\야구>
1	c	광주	<스포츠\야구>
1	a	이승엽	<스포츠\야구>
1	a	baseball	<스포츠\야구>
1	a	홈인원	<스포츠\골프>
1	b	PGA	<스포츠\골프>
1	a	T 샷	<스포츠\골프>

[표 2 연상어와 안정성 랭크의 예]

3.2.2 안정성 순위의 정의
다음으로 중요한 것은 분야 연상어가 시간 경과에 의해 변화하는 것이다. 예를 들면<야구>이면 "투수", "포수" 등은 변하지 않는 안정된 연상어지만, 고교야구의 우승팀이나 선수명은 단기간으로 변화하는 불안정한 연상어이다. 이와 같이 안정성의 낮은 연상어는 고유명사에 많고 특히 인명의 안정성은 매우 낮으므로 연상어에는 다음과 같은 안정성 순위를 정의한다.

[정의 2] 안정성 순위
안정성(stability) 순위를 높은 순서에 보통명사에 a, 고유명사(인명 이외)에 b, 고유명사(인명)에 c를 할당한다.
[표 2]는 수준 1의 분야 연상어와 안정성 랭크를 나타낸다.

3.3 분야 연상어 규칙
다음으로, 복합어에 대한 분야 연상어의 규칙을 설명한다. 단, 이 분야 규칙은, 향후의 연구과제에서 논의하기로 하고 본 장에서는 충분한 평가와 고찰은 주지 않는 것으로 한다.
앞의 예와 같이, "전주 장소"는 <태권도>의 수준 1의 연상어지만, [지명]인 "전주"를 "부산"으로 옮겨놓은 "부산 장소"도 <태권도>를 연상 할 수 있는 것과 같이 [국명]인 "캐나다" 나 "영국"에 옮겨 놓아도 <태권도>의 연상이 가능하다.



[그림2 분야 연상어 규칙의 예]

단, "장소"의 하위어인 "천"에 옮겨놓으면, "전주천"이 되어 <태권도>의 연상 정보는 완전히 사라져 버린다. 따라서, [지명, 국명] "장소"에는 [그림2] 같은 분야 연상어 규칙을 정의 할 수 있다.

이와 같이, 분야 연상어 규칙을 검출하여, 사전에 등록하는 분야 연상어의 총수를 경감시키는 것이 중요하다.

4. 분야 연상어를 이용한 패시지 검색
4.1 시스템의 개요
패시지 검색 시스템의 개요는 그림 3과 같다.

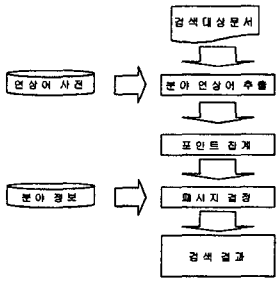


그림 3 시스템 개요

분야 연상어 추출에서는 검색 대상이 되는 문서로부터 한 문장마다의 연상어를 추출한다. 포인트 집계에서는, 분야 연상어 추출로 얻을 수 있던 한 문장마다 연상어를 이용하여 각 분야에 대한 포인트를 산출한다.
패시지 결정에서는 얻을 수 있었던 각 분야의 포인트를 이용하여, 계속도, 전환도를 요구해 그것 보다 화제의 출현, 계속, 전환을 실시하는 것에 의해, 검색 대상이 되는 문서의 패시지를 요구한다.
분야 연상어를 이용하여, 패시지 결정을 실시하기 때문에 사전 처리로서 패시지 검색의 대상이 되는 문서로부터 분야 연상어를 추출할 필요가 있다. 본 방법에서는 AC법에 이용하여 분야 연상어로부터 연상어 사전에 등록된 키워드와 문장과의 캐릭터 라인 조합을 실시하는 것으로 분야 연상어의 추출을 실현한다.

AC법은 단일 패스에서 텍스트 스트링에 포함된 모든 키워드의 위치를 한번에 결정할 수 있는 알고리즘이다.

4.2 패시지의 취득 방법
문서 중 화제의 흐름의 특징을 보면 일련의 화제에는 계속

성이 있다 즉 화제는 산발적이지 않다는 것이다. 또한 화제 흐름은 전환점이 있다. 화제는 중복되지 않는다는 특징을 가진다.
이와 같은 흐름을 고려하여 본 방법에서는 계속도 α를 이용하여 각 분야 마다 계속성을 검색하게 된다.

순서 1 : $\alpha_i = \alpha_{i-1} + \rho \times$ 쇠퇴율
순서 2 : $\alpha_i = \alpha_i + Freq(S_i, F_k)$
단, $Freq(S_i, F_k)$는 Si 문장내의 분야 Fk의 연상어 포인트
ρ는 감쇠율의 영향을 조정한는 파라미터

쇠퇴율이란 화제가 다음 문장으로 이전될 때 쇠퇴한 척도를 말한다.

$$\text{쇠퇴율} = -1 \times \left\{ \frac{\sum_{Sk \in C_{i-1}} Freq(Sk, Fk) + Freq(Si, Fk)}{\text{문서집합} Ci-1\text{의요소수}} \right\}$$

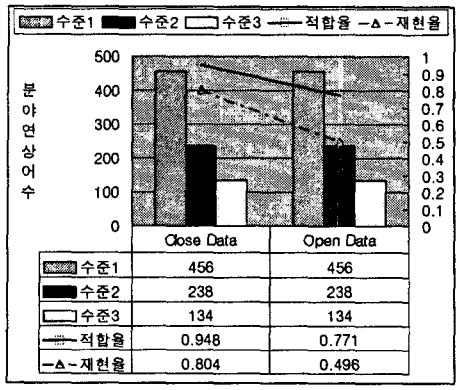
단, C_{i-1}은 S_{i-1} 문장으로부터 거슬러 올라가 분야 연상어가 연속 출현한 문장집합

5. 실험평가
본 시스템을 실험하기 위한 분야체계는 중간 분야50개와 종단 분야 393개로 하였으며 대상 분야는 깊이가 1인 중간 분야 10개에 종단분야 295개로 한정하였다. 문서 데이터는 일간 신문과 CD-ROM데이터, 인터넷에서 수집하였다.
데이터 수집 결과 분야 연상어는 3102개로 평가용 데이터는 각 종단 분야에서 연속한 N(5,10,15,20)행을 추출하여 랜덤하게 5분야를 정리한 파일 30개를 작성하였다.
평가 기준은 적합율과 재현율인데 적합율은 출력결과와 정해진 패시지의 일치되는 문자를 출력 결과의 문자수로 나눈 것이고 재현율은 출력결과와 정해진 패시지의 일치된 문자수를 정해진 패시지의 문자수로 나눈 것이다.

$$\text{적합율} = \frac{\text{출력결과와정답패시지의일치되는문자수}}{\text{출력결과의문자수}}$$

$$\text{재현율} = \frac{\text{출력결과와정답패시지의일치된문자수}}{\text{정답패시지의문자수}}$$

6. 향후 과제 및 결론
본 논문에서는, 패시지를 있는 분야의 화제에 관하여 쓰여진 정리로서 파악하고, 분야 연상어를 이용하고, 패시지를 결정한 방법을 제안하였다.
그리고, 전문가법에 의해 결정된 패시지와, 미리 사람에 의하고 결정한 패시지가 어느 정도 일치하고 있든 지를 비교하고, 제안 방법의 유효성을 평가했다. 그 결과, 50개 문서의 정답 패시지에 대한 평균 적합율이 약80%, 평균 재현율이 약67%로 되어, 전문가법의 유효성이 확인할 수 있었다.
향후 과제로는 현재 수작업으로 이루어지는 분야 연상어의 포인트, 파라미터 ρ의 전 자료 문서로부터의 자동 설정하는 것이 큰 과제이다. 그리고 분야를 스포츠에 국한하여 실험하였으나 보다 폭 넓은 분야와 비교 평가 할 계획이며, 본 방법을 처음에 제시한 여러 가지 방법들과 비교 평가할 수 있는 실험을 계속하고자 한다.



	Close Data	Open Data
수준1	456	456
수준2	238	238
수준3	134	134
적합율	0.948	0.771
재현율	0.804	0.496

그림 4 스포츠에 대한 Open data와 Close data에 대한 분야연상어수와 정밀도

참고문헌
[1] Salton et al. : "Approaches to passage retrieval in full text information systems", In Proc. of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval, pp.49-56, 1993.

[2] Wilkinson : "Effective retrieval of structured documents", In Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval, pp.311-317, 1994.

[3] Mittendorf and Schauble : "Document and Passage Retrieval Based on Hidden Markov Models", In Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval, pp.302-310, 1994.

[4] Callen : "Passage-Level Evidence in Document Retrieval", In Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval, pp.302-310, 1994.

[5] 望月 源, 岩山 真, 奥村 学 : 語彙的連鎖に基づくパッセージ検索, 自然言語処理, Vol. 6, No. 3, pp.101-126, 1999.

[6] 岩山 真, 徳永 健伸 : 確率モデルに基づくパッセージ分類とその応用, 自然言語処理, Vol. 6, No. 3, pp.181-198, 1999.

[7] 水野 浩之, 黄瀬 浩一, 松本 啓之亮 : 単語の出現密度分布と偏出度を用いた図表と説明テキストの対応付け, 情報処理学会論文誌, Vol. 40, No. 12, pp.4400-4403, 1999.

[8] 辻 孝了, 泓田 正雄, 森田 和宏, 青江 順一 : 複合語の分野連想語の効率的決定法, 自然言語処理, Vol. 7, No. 2, 2000.