제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

컴포넌트 검색을 지원하는 시소러스에 의한 질의평가

김귀정*
*건양대학교 IT학부
e-mail:gjkim@konyang.ac.kr

Query Evaluation by Thesaurus to Support Component Retrieval

Gui-Jug Kim*
*Division of Information Technology, KonYang University

요 약

본 논문은 사용자 질의가 가지는 특정 클래스로부터 개념적으로 서로 연관있는 컴포넌트를 검색하기 위하여 퍼지 시소러스를 통한 질의 평가 방법을 이용하였다. 시소러스에 의한 사용자 질의 확장과정은 용어 불일치 문제를 해결함으로써 검색에 대한 일정한 정확도를 보장하면서 재현율을 향상시킬 수 있게 한다. 질의 확장과정의 효율성을 평가하기 위하여 시뮬레이션을 통한 최적의 검색 효율을 나타내는 임계치를 설정하고 재현율과 정확도를 비교하였다.

1. 서론

소프트웨어의 재사용을 위한 효율적인 검색을 위해서는 사용자의 요구정보와 검색하고자하는 컴포넌트의 색인정보를 명확히 표현해야 하며, 정보요구를 만족시키는 적절한 정보들만을 탐색하고 탐색된 정보들에 대해 정보요구의 만족도에 따라 적합성을 부여하는 방법이 필요하다[1]. 이를 위해 시소러스와 같은 지식베이스를 이용한 정보검색 방법들이 많이 제한되었으나, 이처럼 시소러스를 사용하는 일은 많은 검색 시간을 필요로 하며, 심지어는 검색 과정에서 원하는 정보를 찾지 못하는 경우도 발생한다. 또한 질의와 시소러스 용어 사이의 불일치 문제가 발생하여 검색의 재현율을 감소시키는 주된 원인으로 작용하기 때문에 사용자의 효과적인 검색을 저해하게 된다[1].

본 논문에서는 이 단점을 해결하기 위해 퍼지 시소러스를 통한 질의 평가 방법을 이용하였다. 이 검색은 퍼지 시소러스를 이용하여 용어 불일치 문제를 해결함으로써 검색에 대한 일정한 정확도를 보장하면서 재현율을 향상시킬 수 있게 한다. 즉, 컴포넌트와 클래스 사이의 관계를 퍼지 정도로 표현한 시소러스를 이용함으로써 사용자 질의와 정확히 일치하는 컴포넌트뿐아니라 개념적으로 서로 연관된 유사한 의미를 가지는 컴포넌트까지 검색할 수 있게 한다. 본 연구에서는 이를 위해 시소러스에 의한 사용자 질의 확장과정을 거쳤으며, 효과적인 질의 확장과 검색 노이즈의 감소를 위하여 시뮬레이션을 통한 최적의 검색 효율을 나타내는 임계치를 설정하였다.

2. 관련 연구

전통적인 정보 검색은 사용자 질의에 기술된 키워드로 색인된 문서들을 데이터베이스로 검색하고, 이들을 질의와의 관련정도에 따라 제시한다. 검색 모델은 사용자 질의와 문서사이의 관련 정도를 평가하는 방법에 따라 벡터 모델[2], 확률모델[3], 그리고 불리언 모델[4]로 구분된다. 이중 불리언 모델은 매우 단순한 구조를 가지며, 사용자 질의에서 탐색어들 사이의 논리적 관계를 AND나 OR로 비교적 자연스럽게 표현할 수 있기 때문에 다른 모델에 비해 사용자의 요구를 정확히 반영한다는 장점을 가진다. 그러나 컴포넌트와 질의 사이의 관련정도를 평가할 수 없다는 단점도 가지고 있다. 이 단점을 해결하기 위하여 퍼지 멤버쉽 함수에 따라 관련 정도를 평가할 수 있는 퍼지 불리언 모델과 가중치를 갖는 컴포

넌트와 질의간의 유사도를 계산할 수 있는 확장된 불리언 모델이 제안되었다. 퍼지 불리언 모델은 도메인 개념 사이의 관계를 정의한 시소러스와 쉽게 통합될 수 있다는 장점을 가지고 있다. 이 시소러스는 도메인에 매우 의존적이며 문헌과 같이 긴 텍스트 검색에 유용한 특징을 가진다.

3. 컴포넌트 검색을 위한 질의 확장

3.1 퍼지 불리언 질의

본 연구에서는 퍼지 불리언 형태의 질의를 사용하여 각각의 질의어들에 대해 의미적 중요성을 차등있게 표현할 수 있도록 하였다. 퍼지 불리언 모델은 사용자 의도에 따라 질의어 관련 정도를 부여할 수 있을 뿐 아니라, 도메인 개념들 사이의 관계를 정의한 시소러스와 쉽게 통합될 수 있다는 장점이 있다.

다음은 질의를 형성하는 질의어들에 대한 불리언 연산을 표현한 식이다. AND와 OR는 불리언 연산자이며, c는 하나의 질의어이고, α는 사용자가 질의 형성 시 입력한 퍼지 질의어 중요도이다. 포괄적인 의미로 질의를 표현하기 위해서는 퍼지 질의어 중요도인 α를 생략할 수 있는데, 이 경우에는 α=1.0으로 간주된다.

$$Q=(AND|OR)[c_i:\alpha_i]_{i=1}^{n}, \qquad 0\le\alpha\le1$$

퍼지 불리언 질의를 3 가지 형태의 질의로 표준화할 수 있다. 단순질의, 분리질의, 그리고 결합질의가 그것인데, 단순질의는 질의로 하나의 질의어만이 사용된 경우이며, 두 개 이상의 질의어가 있을 경우 분리질의는 OR의 역할을 하고 결합질의는 AND의 역할을 수행한다. 다음은 3 가지 형태의 질의를 정의한 것이다.

단순질의 : $q_\ell=[c_1:\alpha_1]$

분리질의 : $q_{\ell(EXP)}=OR[q]_{i=1}^{n}$

결합질의 : $Q=AND[q_{\ell(EXP)}]_{\ell=1}^{m}$

본 연구에서는 분리질의를 단순질의에 대한 질의 확장의 결과로 정의하였다. 즉, 하나의 단순질의에 대해서 확장된 질의는 모두 OR로 표현된다. 또한, 분리질의로 표현된 확장된 질의어들은 AND 연산을 함으로써 결합질의로 표현될 수 있다. 따라서, 모든 퍼지 불리언 질의는 단순질의의 분리질의에 대한 결합질의로 나타낼 수 있다.

3.2 퍼지 시소러스에 의한 질의 확장과 검색

본 연구의 컴포넌트 검색은 퍼지 불리언 형태로 표현된 사용자 질의를 시소러스를 통해 확장함으로써 이루어진다. 먼저, 한 컴포넌트(i)에 대한 컴포넌트 색인 집합(Collection(i))은 클래스명으로 이루어져 있으며, 각 색인어는 컴포넌트에 대한 가중치를 가지고 있다. 컴포넌트 i에 대한 색인어 c의 가중치를 $W_{com(i,c)}$= α와 같이 표현하기로 한다. 또한 사용자는 단순질의 q=[c:β]를 질의로 입력함으로써 질의어 중요도를 선택할 수 있다. 이때 컴포넌트 i가 사용자 질의 q를 만족하는 정도를 γ라 할때, 이 값은 γ=min(α,β)로 계산된다. 예를 들어 'CommonSocket' 컴포넌트가 'ToolBar : 0.65', 'Document : 0.60', 'Socket : 0.9', 'SocketFile : 0.91', 'DC : 0.87'와 같은 색인 집합을 가지고 있고, 사용자 질의 q가 [SocketFile : 0.9]와 같이 주워졌을 때, 'CommonSocket' 컴포넌트는 질의 q를 0.9(min(0.91, 0.90)) 정도로 만족하고 있음을 뜻한다.

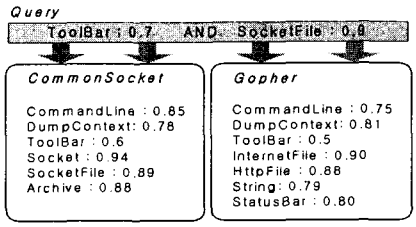

그림 1. 질의 확장과 컴포넌트 검색

그림 1은 사용자 질의가 'ToolBar : 0.7 AND SocketFile : 0.9' 일 때 두 컴포넌트 'CommonSocket'와 'Gopher'의 질의 확장 과정과 컴포넌트 검색 과정을 보여주기 위한 예이다. 'CommonSocket' 컴포넌트의 색인 집합 $Collection_{(CommonSocket)}$은 "CommandLine : 0.85, DumpContext: 0.78, ToolBar : 0.6, Socket : 0.94, SocketFile : 0.89, Archive : 0.88" 이고, 'Gopher' 컴포넌트의 색인 집합 $Collection_{(Gopher)}$는 "CommandLine : 0.75, DumpContext: 0.81, ToolBar : 0.5, InternetFile : 0.90, HttpFile : 0.88, String: 0.79, StatusBar : 0.80"로 정의되어 있다.

먼저, 'CommonSocket' 컴포넌트에 대한 질의 평가와 검색 여부 평가 방법은 다음과 같다. 사용자 질의가 q1=[ToolBar : 0.7] 이고 q2=[SocketFile : 0.9] 라면, 'CommonSocket' 컴포넌트는 질의어 q1을 0.6(min(0.6, 0.7)) 정도로 만족하고 있으며 질의어 q2를 0.89(min(0.89, 0.9)) 정도로 만족하고 있다. 따라서, 'CommonSocket' 컴포넌트가 질의어 q1과 q2를 동시에 만족하는 정도는 min(0.6, 0.89)로 해석되므로, 질의에 대한 컴포넌트의 만족 정도는 0.6으로

평가될 수 있다. 질의에 대한 컴포넌트 만족도가 0.5 이상이면, 사용자 질의에 대해 컴포넌트가 어느 정도 연관성이 높다고 인정되므로 'CommonSocket' 컴포넌트를 검색될 후보 컴포넌트에 포함시킨다.

그러나 'Gopher' 컴포넌트와 같은 경우에는 의미적으로 사용자 질의와 상당히 관련이 있지만, 색인 집합 Collection(Gopher)에 'SocketFile'이 포함되어 있지 않기 때문에 'Gopher' 컴포넌트는 검색되지 않는다. 이 단점은 검색의 재현율을 감소시키는 주된 원인으로 작용된다.

표 1. 퍼지 시소러스 유의어 테이블

class명 / class명	..	InternetFile	HttpFile	Archivet	Menu	..
...	..	...	...	...	...	..
SocketFile	..	0.79	0.73	0.65	0.4	..
...	..	...	...	...	...	..

이 단점을 해결하기 위한 'Gopher' 컴포넌트 검색을 위한 질의 확장 과정은 다음과 같다. 퍼지 시소러스에 의한 유의어 테이블에 의해 "q2=[SocketFile : 0.9]"는 표 1과 같이 질의 확장될 수 있다. 'SocketFile' 클래스는 모든 클래스에 대한 유의값을 가지고 있으며, 이중 유의값이 0.7이상에 해당하는 클래스들만이 질의 확장의 대상이 된다. 이는 정확도를 적절히 유지하면서도 재현율이 높게 나타나는 임계치 범위를 시뮬레이션 통하여 설정한 것으로, 4. 성능평가에 자세히 설명하고 있다. 이에 따라 퍼지 시소러스에서 질의어 q2='SocketFile' 클래스의 질의 확장 집합 Exp(SocketFile)은 {SocketFile : 1.0, InternetFile : 0.79, HttpFile : 0.73}과 같이 구성될 수 있다. 여기에서 질의어 q2의 'SocketFile' 중요도가 사용자에 의해 0.9로 설정되었기 때문에 질의 확장 집합 Exp(SocketFile)의 각 확장 질의에 대한 유의값이 조절되어야 하는데, 질의어에 주워진 사용자 중요도와 각 확장 질의의 유의값 중 적은 값을 선택한다. 그러므로 질의어 q2=[SocketFile : 0.9]에 대한 최종적인 질의 확장 집합 Exp(SocketFile)은 {SocketFile : 0.9, InternetFile : 0.79, HttpFile : 0.73}과 같이 확장된다.

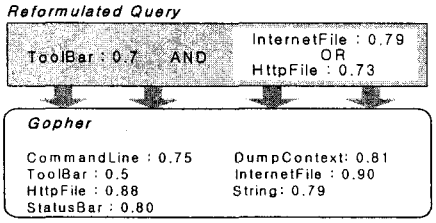

그림 2. 확장된 질의

그림 2는 이를 바탕으로 재형성된 질의를 나타낸다. 질의가 q1=[ToolBar : 0.7] 이고 q2=[InternetFile : 0.79 OR HttpFile : 0.73]이므로, 'Gopher'는 질의어 q1을 0.5(min(0.5, 0.7)) 정도로 만족하고 있다. 또한 질의 q2에 대해 확장된 각 질의어의 질의 중요도와 컴포넌트 가중치 중 적은 값을 선택한 후 퍼지 OR 연산을 수행하여 0.79의 값을 얻을 수 있다. 이는 'Gopher'가 질의어 q2를 0.79 정도로 만족하고 있음을 나타낸다. 따라서, 'Gopher' 컴포넌트가 질의어 q1과 q2를 동시에 만족하는 정도는 min(0.5, 0.79)로 해석되므로, 질의에 대한 컴포넌트의 만족 정도는 0.5로 평가되어 'Gopher' 컴포넌트는 후보 컴포넌트에 포함되어 검색되어질 수 있다.

4. 성능평가

시뮬레이션에 사용된 컴포넌트는 Visual C++ Class Library로 구성하였다. 본 연구에서 사용된 총 131개의 컴포넌트는 11개의 Class Concept Category(CCC)에 포함되어 있으며, 각 CCC는 컴포넌트에 따라 최소 1개에서 최대 22개까지의 클래스로 구성되어 있다. 컴포넌트에 나타난 CCC의 평균 클래스 수는 약 8개이고, 모든 CCC에는 중복을 포함하여 총 1023개의 클래스가 존재하며 독립된 303개의 클래스로 구성되어 있다.

4.1질의 확장 임계치 평가

본 연구에서는 컴포넌트 검색의 재현율을 최대한 보장해줄 수 있는 최적의 질의 확장 임계치를 시뮬레이션을 통하여 설정하였다. 이를 위해 시소러스 내에서 임의의 클래스 10개를 선택하여 클래스에 대한 유사 확장 집합(similar expended sets)을 선정하였다. 10개의 클래스에 대해 확장된 질의와 유사 확장 집합과의 비교를 통하여 정확도와 재현율을 측정하였다. 이때 질의 확장에 있어 유의값을 0.6에서부터 1.0까지 0.05의 간격으로 각각의 정확도와 재현율을 측정하였다. 정확도와 재현율을 측정하는 방법은 다음과 같다[5].

$$\text{재현률} = \frac{\text{확장된 유의어 중 유사확장 집합에 속한 유의어의 수}}{\text{유사확장 집합의 수}}$$

$$\text{정확도} = \frac{\text{확장된 유의어 중 유사확장 집합에 속한 유의어의 수}}{\text{확장된 유의어의 수}}$$

그림 3은 정확도와 재현율의 평균을 임계값에 따라 그래프로 보여준다. 임계값이 높아질수록 정확도가 좋아지고, 낮아질수록 재현율이 향상됨을 알 수 있다. 그러나 임계값 0.7 미만이 되면 정확도가 0에 가깝게 되어 검색 효율이 떨어지고, 임계값 0.8부터

는 0.7과 정확도에 있어서는 별 차이가 나지 않지만 재현율은 상당히 떨어짐을 알 수가 있다. 따라서 본 연구에서는 정확도를 유지하면서 재현율을 최대한 보장할 수 있는 범위인 임계값 0.7 이상을 질의 확장의 범위로 설정하였다.

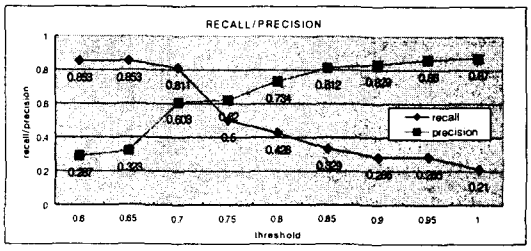

그림 3. 임계값에 따른 정확도/재현율 비교

4.2 질의 확장에 따른 성능 평가

임계값 0.7이상의 질의 확장과 이에 따른 시소러스를 성능 평가하였다. 방법은 하나의 질의에 대하여 5번의 확장을 시행한 후, 정확도와 재현율의 변화를 비교하는 것이다. 임계값 0.7 이상인 모든 클래스를 질의에 포함시키고, 새롭게 질의에 포함된 클래스에 대해서 다시 0.7 이상인 클래스를 질의에 포함시키는 방법으로 총 5번의 질의 확장을 시행한다. 이는 확장 과정에 따라 질의에 대한 재현율과 정확도의 변화를 알기 위함이다. 표 2에서처럼 "Window"를 5번 확장한 결과 모두 8 개의 질의로 확장되었다. 10개의 클래스에 대해 시행한 결과의 정확도와 재현율 비교가 표 2와 그림 4에 나타난다. 확장이 진행될수록 정확도가 떨어지고, 재현율이 향상됨을 알 수 있다. 그러나 확장이 한번도 이루어지지 않은 경우에는 정확도는 높지만, 재현율이 낮아 검색 효율이 떨어지며, 확장이 3번 이상 이루어진 경우에는 재현율에 있어서는 별 차이가 나지 않지만 정확도가 상당히 떨어짐을 알 수가 있다.

표 2. "Window" 클래스 확장 과정

확장 과정	질의
Q1	**Window**
Q2	Window **View : 0.84** **AnimateCtrl : 0.79**
Q3	Window View **FrameView : 0.93** **ScrollView : 0.77** **CtrlView : 0.85** AnimateCtrl
Q4	Window View FrameView CtrlView ScrollView **RecordView : 0.80** AnimateCtrl
Q5	Window View FrameView CtrlView ScrollView RecordView **DaoRecordView : 0.92** AnimateCtrl

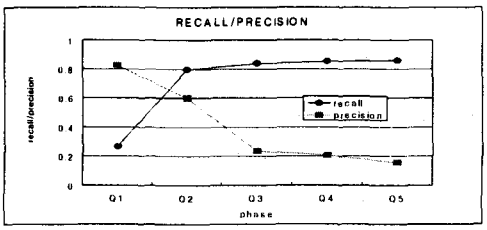

그림 4. 확장에 따른 정확도/재현율 비교

5. 결론

본 논문은 사용자 질의가 가지는 특정 클래스로부터 개념적으로 서로 연관있는 컴포넌트를 검색하기 위하여 퍼지 시소러스를 통한 질의 평가 방법을 이용하였다. 이 검색은 사용자 질의와 정확히 일치하는 컴포넌트뿐아니라 개념적으로 서로 유사한 의미를 가지는 컴포넌트까지 검색할 수 있게하여 용어 불일치 문제를 해결 함으로써 검색에 대한 일정한 정확도를 보장하면서 재현율을 향상시킬 수 있었다. 이를 위해 시뮬레이션을 통해 질의확장 범위를 임계값 0.7이상으로 설정하였으며, 질의확장 횟수는 한번으로 설정하여 최적의 검색효율을 얻을 수 있었다.

참 고 문 헌

[1] H.J.Peat and P.Willett, "The Limitation of Term Co-occurrence Data for Query Expansion in Document Retrieval System", Journal of the American Sociery for Onfprmation Science, Vol.42, No.5, pp. 378-383, 1991.

[2] G. Salton and M. J. McGill "Introduction to Modern Information Retrieval," McGraw-Hill Information Editor, 1987.

[3] N. Fuhr, "Probabilistic Models in Information Retrieval," The Computer Journal, Vol. 35, No. 3, pp. 243-255, 1992.

[4] B. Y. Ricardo and R. N. Berthier, "Modern Information Retrieval," Addison-Wesley, 2000.

[5] E. Damiani, M. G. Fugini and C. Bellettini, "Aware Approach to Faceted Classification of Object-Oriented Component,"ACM Transaction on Software Engineering and Methodology, Vol.8, No.4, pp.425-472. Oct. 1999.