논문 2008-1-8

퍼지 지식베이스를 이용한 효과적인 다언어 문서 검색

Effective Cross-Lingual Text Retrieval using a Fuzzy Knowledge Base

최명복*

Myeong-Bok Choi*

요 약

다언어 문서검색(CLTR; Cross-Lingual Text Retrieval)은 하나의 언어로 질의가 주어질 때, 그 질의의 언어와는 다른 언어로 되어 있는 문서들을 검색하는 정보 검색을 말한다. 본 논문에서는 두 언어 사이의 용어들 간에 부분 매칭을 다룰 수 있도록 하기 위해 퍼지 다언어 시소러스 기반의 다언어 문서검색 시스템을 제안한다. 제안된 다언어 문서검색 시스템에서는 효과적인 추론을 위해 퍼지 용어 매트릭스를 정의하여 이용한다. 정의된 퍼지 용어 매트릭스에서 용어들 간의 모든 관련도가 전이폐쇄 알고리즘을 이용하여 추론함으로써 용어들 간의 묵시적인 링크가 모두 검색에 반영된다. 이에 따라 제안된 방법은 인간 전문가에 좀 더 가까운 정보검색을 수행하여 검색 효과를 높이게 된다.

Abstract

Cross-lingual text retrieval(CLTR) is the information retrieval in which a user tries to search a set of documents written in one language for a query another language. This thesis proposes a CLTR system based on fuzzy multilingual thesaurus to handle a partial matching between terms of two different languages. The proposed CLTR system uses a fuzzy term matrix defined in our thesis to perform the information retrieval effectively. In the defined fuzzy term matrix, all relation degrees between terms are inferred from using the transitive closure algorithm to reflect all implicit links between terms into processing of the information retrieval. With this framework, the CLTR system proposed in our thesis enhances the retrieval effectiveness because it is able to emulate a human expert's decision making well in CLTR.

Key Words : Cross-lingual Text Retrieval, Multilingual Information Retrieval, Thesaurus, Information Retrieval, Knowledge Base

Ⅰ. 서 론

정보검색 시스템의 주요한 목적을 극대화하기 위해서는 첫째 사용자의 정보요구와 문서의 색인인 검색형태(Search Patterns)를 명확하게 표현하는 것이고[1], 둘째는 정보요구를 만족시키는 적절한 정보들만을 탐색하고 탐색된 정보들에 대해 정보요구의 만족도에 따른 적합성 순위를 부여하는 것이다. 이러한 정보검색 시스템의 주요한 목적의 극대

*정회원, 강릉대학교 컴퓨터공학부

접수일자: 2008.1.20, 수정완료일자: 2008.2.12

화는 질의의 언어와 검색 대상 문서의 언어가 동일한 단일언어 정보검색(MLIR; Mono-Lingual Information Retrieval)에서 뿐만 아니라 질의의 언어와 검색 대상 문서의 언어가 다른 언어로 쓰여진 문서를 검색하는 다언어 정보검색(CLIR; Cross-Lingual Information Retrieval)에서도 마찬가지이다.

인터넷의 급속한 발전과 대중화에 따라 자국 언어 외에 다른 언어로 쓰여진 웹 문서들이 급격히 증가하면서 다언어 정보검색에 대한 연구가 활발히 진행되고 있다[2-6].

다언어 정보검색에는 어휘 불일치(Vocabulary Mismatch)와 다언어간 의미적 부정확성(Semantic Imprecision)이라는 2가지의 주요한 문제점들이 존재한다[4]. 어휘 불일치는 질의 언어와 검색 대상 문서의 언어가 서로 다르기 때문에 기본적으로 두 언어 사이에서 나타나는 어휘간의 차이점을 의미하며 다언어 정보검색의 핵심문제이다[3]. 또한 언어는 그 나라의 문화를 반영하고 있기 때문에 하나의 언어에서의 개념은 다른 언어에서 사전적으로 정확하게 일치하는 개념이 존재하지 않을 수 있지만, 의미가 약간 다른 개념들이 존재할 수 있다. 이러한 종류의 다언어간 의미적 부정확성은 어휘 불일치 문제를 더욱 복잡하게 만든다[4]. 이러한 문제점들은 사용자의 정보 요구에 가장 관련 있는 문서들을 검색하는 데에 큰 영향을 주기 때문에 다언어 정보검색에서 이러한 두 언어의 어휘 차이를 극복하는 방법이 주요한 관심사가 아닐 수 없다.

논문[4-5]는 다언어 정보검색에서 발생되는 어휘 불일치와 다언어간 의미적 부정확성 문제들을 해결하기 위하여 퍼지 전문가 시스템 기법을 이용한 다언어 정보검색 시스템을 제안하였다. 이 시스템에서는 인간 전문가와 비슷한 방법으로 검색을 수행할 수 있도록 하기 위해서 용어들 간의 부분적 의미 관계를 표현할 수 있는 지식 베이스(Knowledge Base)인 퍼지 다언어 시소러스 기반의 퍼지 근사 추론(Approximate Reasoning)을 수행하여 사용자의 질의에 대한 문서 검색의 효과를 높이게 된다.

그러나 논문[4-5]에 있는 퍼지 다언어 전문가 시스템의 추론 엔진에서 사용하고 있는 근사 추론은 퍼지 다언어 시소러스에 있는 묵시적인 링크(Implicit Link)를 무시하고 추론하게 된다. 용어들 간의 묵시적인 링크를 무시한다는 것은 지식 베이스에 표현된 용어들 간의 모든 의미적 관련성을 효과적으로 사용하지 못한다는 의미이다. 이렇게 되면 사용자의 정보 요구에 맞는 적절한 문서의 검색 결과가 인간 전문가의 검색 결과와 매우 다를 수 있으며, 때에 따라서는 검색된 문서들의 순위를 결정하는데 어려울 수 있다.

본 논문에서는 이러한 논문[4-5]의 문제점을 해결하기 위하여 퍼지 다언어 시소러스와 같은 지식 베이스에서 용어들 간의 묵시적인 링크를 찾기 위한 전이 폐쇄 알고리즘[7]을 적용하는 방법을 제안한다. 본 논문에서 제안된 방법은 논문[4-5]의 문제점을 해결함으로써 좀 더 인간 전문가에 가까운 정보 검색으로 검색 효과를 높일 수 있게 한다.

본 논문의 나머지는 다음과 같이 구성된다. 2장에서는 논문[4-5]에서 제안된 다언어 정보검색 시스템을 살펴보고 문제점을 논한다. 3장에서는 용어 매트릭스를 이용한 다언어 정보검색 방법을 제안한다. 4장에서는 결론 및 향후 연구 방향을 제시한다.

II. 다언어 정보 검색 시스템

이 부분에서는 논문[4-5]에서 제안한 다언

어 정보 검색 시스템의 구조와 추론 방법을 살펴보고 하나의 예제를 통해 다언어 정보 검색 시스템에서 사용하고 있는 퍼지 다언어 시소러스를 이용한 추론 방법의 문제점을 살펴본다.

2.1 퍼지 다언어 전문가 시스템의 추론

퍼지 다언어 전문가 시스템(Fuzzy Multilingual Expert System)[4-5]은 질의 처리기(Query Processor), 퍼지 다언어 시소러스(Fuzzy Multilingual Thesaurus), 다언어 문서 데이터베이스(Multilingual Document Database) 그리고 추론엔진(Inference Engine)으로 구성된다. 질의 처리기는 사용자의 정보 요구를 분석하는 프런트-엔드(Front-End) 모듈이며 다언어 문서 데이터베이스는 문서에 대한 유일한 식별자들과 그 문서들을 기록한 언어와 동일한 언어로 된 색인 용어들을 포함하고 있는 관계형 데이터베이스이다. 지식베이스인 퍼지 다언어 시소러스는 다언어 키워드들과 다언어 키워들 간의 의미적 관계들로 구성된 정보 구조(Information Structure)로 다음과 같은 명제들로 표현되는 사실(facts)들의 집합이다.

"용어 a와 b는 의미상으로 서로 다언어 관계를 가지고 있다."

여기서 용어 a와 용어 b는 서로 다른 언어로 쓰여진 용어들이다. 논문[4-5]에서는 이러한 언어적 관계를 2항 퍼지 관계(Fuzzy Relation)[7]을 도입하여 다음과 같은 식으로 표기하였다. 즉, 퍼지 다언어 시소러스인 FT의 개념은 두 용어들의 집합 A와 B가 있을 때 집합의 곱인 A×B 내의 2항 퍼지 관계 FT로 정의된다.

$$FT = \{(a, b), \mu_{FT}(a, b) | (a, b) \in A \times B, \mu_{FT}(a, b) \in [0, 1]\} \quad (1)$$

여기서 $\mu_{FT}(a, b)$: A×B → [0, 1]은 FT 내의 순서쌍인 (a, b)의 소속 정도를 결정하는 소속 함수이다. 소속 정도는 관계 FT에 대해 용어 a와 b가 의미적으로 어느 정도 관련 있는지를 나타낸다. 이에 따라 퍼지 다언어 시소러스에서 어떤 하나의 용어는 그 용어와 관련된 관련 정도의 순위에 따라 그 용어와는 다른 언어로 쓰여진 용어들과 1:n으로 대응시킨다.

논문[4-5] 등의 정보검색 시스템에서 추론엔진은 실제로 검색을 수행하는 검색 모듈이다. 검색 모듈은 지식베이스에 표현된 의미적 지식을 사용하여 사용자의 질의와 각 문서 사이의 관련 정도를 추론하게 된다. 이러한 관련 정도를 문서에 대한 RSV(Retrieval Status Value)라고 부른다.

인간 전문가와 비슷한 방법, 즉 용어들 간의 부분적 의미 관계를 표현할 수 있는 퍼지 다언어 시소러스 기반의 근사 추론을 수행할 수 있도록 퍼지 다언어 전문가 시스템의 추론엔진[4-5]은 퍼지 다언어 시소러스에 의해 제공되는 다언어의 의미적 관계를 이용하는 퍼지 추론을 수행함으로써 문서 검색의 효과를 높이게 된다. 퍼지 추론은 합성추론규칙(Compositional Rule of Inference)[7]을 통하여 증상들을 조합함으로써 퍼지 IF-THEN 규칙들의 집합으로부터 결론을 이끌어 내게 된다. 또한 IF-THEN 규칙들에 의해 추론된 각각의 결론들은 하나로 합성된다. 즉, 퍼지 추론 기법에 따라 다언어 정보 검색 시스템은 다음과 같이 퍼지 IF-THEN 규칙들에 의해 표현되어 각각 퍼지 관계로 대응시키게 된다.

명제 퍼지 관계
IF D is indexed by $d \in L_D \Rightarrow IND(D, L_D)$
AND $d \in L_D$ is related to $d' \in L_Q \Rightarrow FT(L_D, L_Q)$
AND $d' \in L_Q$ is needed by $Q \Rightarrow NEED(L_Q, Q)$
THEN D is relevant to $Q \Rightarrow REL(D, Q)$

여기서 $IND(D, L_D)$는 문서 D와 그 문서의 색인어를 관련 정도에 따라 대응시켜 주는 색인 함수이다. $FT(L_D, L_Q)$는 (1)에서 정의된 퍼지 다언어 시소러스이다. $NEED(L_Q, Q)$는 사용자의 질의 Q와 질의 용어를 관련시키는 퍼지 질의 표현 함수이다. REL(D, Q)는 특정의 질의 Q에 대해 각각의 문서가 어느 정도 관련 있는지를 의미하는 관련 정도를 할당하는 퍼지 매칭 함수이다. 또한 여기서 질의의 용어들의 가중치, 문서에 대한 색인어의 가중치, 퍼지 다언어 시소러스 내의 용어들 간의 관련도, 그리고 질의와 문서의 관련도는 모두 [0, 1] 사이의 값을 가지게 된다.

이와 같은 경우에 REL(D, Q)는 IF-THEN 규칙의 전반부를 표현하고 있는 퍼지 관계에 대한 합성 연산자를 적용하여 다음의 그림과 같이 퍼지 매트릭스로 변환하여 추론하게 된다.

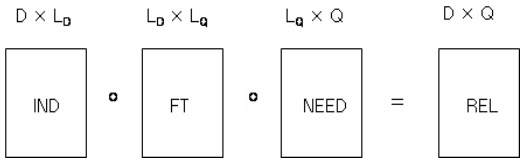

그림 1. 퍼지 관계의 합성에 의한 REL(D, Q)의 추론
Fig. 1 Inference of REL(D, Q) by composition of fuzzy relations

따라서 다음의 수식과 같이 REL(D, Q)는 추론된다.

$$REL = IND \circ FT \circ NEED \qquad (2)$$

여기서 $\mu_{REL}(D, Q) = \mu_{IND \circ FT \circ NEED}(D, Q)$는 문서 D가 질의 Q에 어느 정도 관련되어 있는지를 나타내는 소속 함수이다. 이에 따라 추론엔진은 질의 처리기로부터 질의가 주어질 때, 퍼지 다언어 시소러스의 용어들 간의 관련도를 기반으로 각각의 문서 D의 관련 정도인 RSV ∈ [0, 1]을 (2)의 수식에 의해 계산하게 된다. 얻어진 결과는 RSV의 크기 순으로 정렬하여 사용자에게 반환된다.

2.2 추론의 문제점

이상의 2.1절에서 설명된 것과 같이 퍼지 다언어 전문가 시스템[4-5]은 주로 기존의 다언어 정보 검색 시스템의 문제점들인 어휘 불일치 문제와 다언어 사이의 의미적 부정확성 문제를 해결하기 위한 효과적인 방법을 제안하였다. 즉, 이 방법에서는 사용자의 질의에 있는 용어, 문서에 대한 색인어의 중요성, 그리고 퍼지 다언어 시소러스를 기반으로 하는 퍼지 관계의 합성 추론 규칙을 연산함으로써 각각의 문서에 대한 RSV 값을 추론해 낼 수 있었다.

그러나 퍼지 다언어 전문가 시스템의 추론엔진에서 사용하고 있는 퍼지 매트릭스를 이용한 합성 추론은 퍼지 다언어 시소러스에 있는 묵시적인 링크(Implicit Link)를 무시하고 추론하게 된다. 용어들 간의 묵시적인 링크를 무시하게 되면 실제로는 어떤 문서가 사용자의 정보요구에 관련되어 있지만 관련되지 않은 것으로 추론되어 원하는 문서가 검색되지 않을 수도 있으며, 때에 따라서는 검색된 문서들의 순위를 결정하는데 어려울 수 있다.

예를 들어 다음의 그림2와 같이 3개의 문서와 4개의 용어들로 구성된 예를 살펴보자.

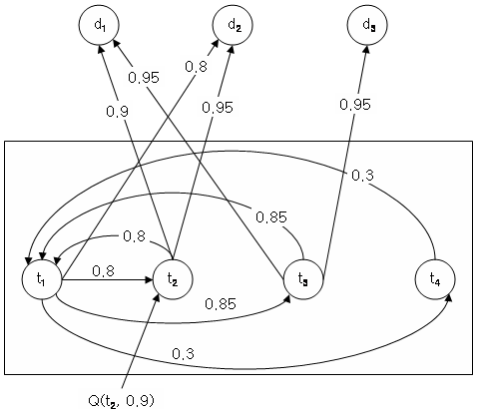

그림 2. 추론을 위한 그래프 표현의 예
Fig 2. An Graph Representation Example for Inference

그림2에서 보는 것과 같이 문서 d_1은 용어 t_2, t_3을 색인어로, 문서 d_2는 용어 t_1, t_2를 색인어로, 그리고 문서 d_3은 용어 t_3을 직접적인 링크에 의한 색인어로 가지고 있다. 다시 말해서 문서 d_1은 용어 t_1 관점에서 볼 때 t_2 또는 t_3을 경유하는 간접적인 링크만 존재하며 직접적으로 d_1과 연결 관계를 가지고 있지는 않다.

그림2와 같이 $Q(t_2, 0.9)$라는 사용자 질의가 주어지면 논문[4-5]에서 제안된 퍼지 다언어 전문가 시스템의 추론 엔진에서는 식(2)을 이용하여 다음과 같이 추론할 것이다.

$$\begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array}\begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 0.00 & 0.90 & 0.95 & 0.00 \\ 0.80 & 0.95 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.95 & 1.00 \end{pmatrix} \end{array} \circ$$

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}\begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 1.00 & 0.80 & 0.85 & 0.30 \\ 0.80 & 1.00 & 0.00 & 0.00 \\ 0.85 & 0.00 & 1.00 & 0.00 \\ 0.30 & 0.00 & 0.00 & 1.00 \end{pmatrix} \end{array} \circ$$

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}\begin{array}{c} Q \\ \begin{pmatrix} 0.00 \\ 0.90 \\ 0.00 \\ 0.00 \end{pmatrix} \end{array} = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array}\begin{array}{c} Q \\ \begin{pmatrix} 0.90 \\ 0.90 \\ 0.00 \end{pmatrix} \end{array}$$

이상의 추론 결과에서 보는 것과 같이 문서 d_1과 d_2는 동등한 순위로 사용자의 질의 $Q(t_2, 0.9)$에 각각 0.9 정도의 관련도를 가지고 있는 문서로 검색되며, 문서 d_3은 0.0의 관련도를 가지므로 검색되지 않게 된다.

그러나 문서 d_3은 실제로 사용자의 질의 $Q(t_2, 0.9)$에 대한 관련 있는 문서로 검색되어야 한다. 왜냐하면 그림 2에서 보는 것과 같이 문서 d_3은 질의에 사용된 용어 t_2로부터 t_1과 t_3을 경유하여 문서 d_3과 관련되어 있기 때문이다. 이와 같이 다언어 전문가 시스템의 추론 엔진[4-5]에서 사용하고 있는 추론 기법에서 사용자의 질의 $Q(t_2, 0.9)$에 관련된 문서로 문서 d_3을 검색하지 못하는 이유는 용어들 간의 묵시적인 링크를 무시하기 때문이다. 다시 말해서 그림2에서 보면 용어 t_2로부터 문서 d_3까지 직접적으로 연결된 링크는 없지만 $t_2 \rightarrow t_1 \rightarrow t_3 \rightarrow d_3$으로 연결되는 묵시적인 링크가 존재한다. 이러한 묵시적인 링크들을 무시함으로서 사용자의 질의 $Q(t_2, 0.9)$에 대해 문서 d_3은 관련되지 않은 문서로 평가되어 검색되지 않게 된다.

실제적으로 각 문서에 대한 색인 용어들이 어느 정도 관련되어 있는 지를 나타내는 색인어들의 가중치는 그 분야의 전문가에 의해 결정될 수 있다. 이 경우 전문가는 어떤 색인 용어들에 대해서는 특정 문서로의 관련성을 그림 2에서와 같이 무시할 수 있다. 무시된 관련성은 그림 2에서와 같이 용어와 용어 사이의 관련성을 나타내는 모든 링크를 찾아봄으로써 추론될 수 있다. 즉, 정보검색 시스템에서 지식베이스로 사용되는 시소러스 내부의 용어들 간의 모든 직·간접적인 관련도의 추론을 사용자의 질의 평가에 반영하면 된다.

III. 지식베이스의 표현과 추론 방법

3.1 지식베이스의 표현

Quillian[8]은 인간의 지식을 표현하는 방법으로 의미 네트워크(Semantic Network)를 제안하였다. 의미 네트워크는 아크(Arcs) 또는 화살표에 의해 명칭 노드(Labeled Node)들을 상호 연결하는 방향성 그래프로 광범위하게 묘사된다. 여기서 노드는 개념을 나타내고 링크는 개념들 사이를 여러 가지 종류의 관계로 연결시킨다.

많은 정보검색 시스템들은 전통적으로 문서에 독립적인 방법으로 특정 주제 문제를 설명하기 위해서 시소러스를 사용하고 있다. 시소러스는 노드와 링크로 구성되며 개념들을 연결하는 다소 제약된 의미 네트워크의 일종으로 볼 수 있다. 논문[9]에서는 시소러스의 구조를 트리(Tree), 가중치 트리(Weighted Tree), 유향 무순환 그래프(Directed Acyclic Graph; DGA), 가중치 DGA, 그리고 그래프(Graph)의 5가지로 분류하였다.

논문[10]은 퍼지 정보검색을 위한 개념 네트워크를 제안하였다. 개념 네트워크는 노드와 유향 링크(Directed Link)로 구성된다. 노드는 개념 또는 문서를 표현하며, 각 링크는 두 개념들을 연결하든지 또는 특정한 개념 C_i와 문서 d_i를 연결한다. 또한 링크에는 [0, 1] 사이의 가중치가 부여된다. 링크에 부여된 가중치 w_{ij}는 개념 t_i와 t_j 사이 또는 개념 t_i와 문서 d_i사이의 관련 정도를 의미하게 된다.

이상에서와 같이 단일 언어(Monolingual) 정보검색을 위한 다양한 그래프 형태의 시소러스를 이용한 검색 방법[9-11]은 용어들 간의 관련성을 추론하기 위해 그래프를 탐색해야 하는데 이는 추론 절차를 매우 느리게 할 수 있다. 특히 사용자의 질의가 복합 질의(AND 또는 OR 연결자로 구성된 질의)로 구성된 경우 더욱 비효율적이다. 이러한 제한점은 결과적으로 실제적인 정보검색 시스템이 구현되었을 때, 대부분의 사용자들에게 만족할 만한 검색 속도를 제공하지 못할 것이다.

이러한 그래프 탐색에 따른 문제점을 해결하기 위하여 논문[12]은 전이 폐쇄 알고리즘[7]을 단일 언어 정보검색을 위한 시소러스에 적용하는 방법을 제안하였다.

3.2 퍼지 용어 매트릭스를 이용한 추론 기법

2.2절에서 살펴본 것과 같이 논문[4-5]에서 제안한 퍼지 다언어 전문가 시스템의 추론 엔진에서는 퍼지 매트릭스를 이용하여 효과적인 추론을 수행하지만 합성 추론에 따른 퍼지 다언어 시소러스에 있는 묵시적인 링크(Implicit Link)를 무시하고 추론함으로써 인간 전문가와는 매우 다른 검색 결과를 제시할 수 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 전이 폐쇄 알고리즘[7]을 퍼지 다언어 시소러스에 적용함으로써 논문[4-5]에서 묵시적인 링크를 무시함으로써 발생하는 문제점을 해결하고자 한다.

본 논문에서는 이를 위해 퍼지 다언어 시소러스를 위한 퍼지 용어 매트릭스를 다음과 같이 정의한다. 정의된 퍼지 용어 매트릭스는 그림 2에서 사각형 내에 있는 그래프 형태의 시소러스를 용이하게 모델링 하게 된다.

[정의3.1] 용어들의 집합 $T_{er} = \{t_1, t_2, t_3, ..., t_n\}$일 때 퍼지 용어 매트릭스 FMT는 다음과 같은 특성을 갖는 퍼지 매트릭스[13]이다. 여기서 FMT(t_i, t_j)는 인접한 용어 t_i로부터 t_j까지의 관련도를 나타낸다. 관련도는 [0, 1] 사이의 값이다.

1) 반사관계
$FMT(t_i, t_i) = 1, \forall t_i \in T_{er}$.
2) 대칭관계가 아닐 수 있다.
$FMT(t_i, t_j) \neq FMT(t_j, t_i)$.
3) 전이관계
$FMT(t_i, t_k) \geq \underset{t_j \in T_{er}}{MaxMin}[FMT(t_i, t_j), FMT(t_j, t_k)]$.

[정의3.2] 퍼지 용어 매트릭스 FMT은 다음과 같이 표현된다.

$$FMT = \begin{pmatrix} f_{11} & f_{12} & ... & f_{1n} \\ f_{21} & f_{22} & ... & f_{2n} \\ . & . & ... & . \\ f_{n1} & f_{n2} & ... & f_{nn} \end{pmatrix}$$

여기서 n은 용어들의 개수이며, f_{ij}는 용어 t_i로부터 용어 t_j 까지의 관련 값(단, $f_{ij} \in [0, 1]$, $1 \leq i \leq n$, $1 \leq j \leq n$)이다.

또한 그림 2과 같은 그래프에서 문서에 대한 색인어의 관계는 다음과 같이 문서 디스크립터로 표현할 수 있다. 여기서 t_i는 문서에 대한 색인어의 집합, n은 T의 원소 수, 그리고 w_i는 그 문서에 대한 색인어 t_i의 가중치를 나타낸다.

$$d = <(t_1, w_1), (t_2, w_2), ... , (t_n, w_n)> \quad (3)$$
(단, $t_i \in T$, $w_i \in [0, 1]$, $1 \leq i \leq n$)

문서 집합 D는 가중치 w에 대해 d와 t간의 문서 디스크립터 벡터로 다음과 같이 표현할 수 있다. 여기서 m은 문서들의 수이고 n은 용어들의 수이다. w_{ij}(단, $w_{ij} \in [0, 1]$, $1 \leq i \leq m$, $1 \leq j \leq n$)는 문서 d_i에 대한 개념 t_j의 가중치를 나타낸다.

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ . \\ d_m \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & ..... & t_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & ..... & w_{1n} \\ w_{21} & w_{22} & w_{23} & ..... & w_{2n} \\ . & . & . & ..... & . \\ w_{m1} & w_{m2} & w_{m3} & ..... & w_{mn} \end{pmatrix} \end{array}$$

문서 디스크립터 매트릭스 D에서 문서 d_i에 대한 용어 t_j의 관련성 정도는 전문가에 의해 결정될 수 있다. 그러나 전문가는 어떤 용어들에 대해서는 특정 문서로의 관련성을 무시할 수 있다. 무시된 관련성은 용어와 용어 사이의 관련성을 나타내는 퍼지 용어 매트릭스의 모든 연결 관계를 찾아봄으로써 구할 수 있다. 이러한 퍼지 용어 매트릭스에서 용어들 간의 모든 연결 관계는 퍼지 용어 매트릭스 FMT의 전이폐쇄(Transitive Closure)에 의해 쉽게 파악할 수 있다. 퍼지 용어 매트릭스 FMT의 전이폐쇄 R_T는 다음과 같은 간단한 알고리즘[7]에 의해 구할 수 있다. 정의에서 ∪는 합집합 기호로 퍼지 집합의 Max 연산자, 그리고 ∘는 Max-Min 합성 연산을 의미한다.

[정의3.3] 관계 R(X, X)가 있을 때, 관계 R(X, X)의 전이폐쇄 R_T(X, X)는 다음과 같은 3단계로 구성되는 간단한 알고리즘에 의해 결정될 수 있다.

1. $R' = R \cup (R \circ R)$.
2. If $R' \neq R$, make $R = R'$ and go to Step 1.
3. Stop: $R' = R_T$.

이상에서 정의한 퍼지 용어 매트릭스 FMT와 문서 디스크립터 벡터는 논문[4-5]에서 제안한 식 (2)의 항목 중 퍼지 다언어 시소러스인 FT와 문서에 대한 색인어를 의미하는 IND 항목을 각각 그대로 모델링 한다. 따라서 논문[4-5]에서 제안한 식 (2)에서 퍼지 다언어 시소러스인 FT 대신에 퍼지 용어 매트릭스 FMT의 전이폐쇄 R_T를 이용하면 논문[4-5]에서 묵시적인 링크를 무시함으로써 발생하는 문제점을 해결할 수 있는 것이다. 이에 따라 식 (2)은 다음과 같은 식 (4)에 의해 대체될 수 있다. 또한 추론 엔진

의 실제 추론은 다음의 식 (5)을 이용하여 연산하게 된다. 식 (5)에서 FMT_R_T는 퍼지 용어 매트릭스 FMT에 전이 폐쇄 알고리즘을 적용한 결과이다.

$$REL = D \circ FMT \circ NEED \quad (4)$$
$$REL = D \circ FMT\_R_T \circ NEED \quad (5)$$

그러면 2.2절의 그림 2를 예로 하여 본 3.2절에서 제안된 수식 (5)의 방법으로 추론을 수행해 보자. 그림 2의 그래프는 식 (4)에 의해 문서 디스크립터 D, 퍼지 용어 매트릭스 FMT, 그리고 사용자의 질의 NEED로 다음과 같이 각각 표현된다.

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 0.00 & 0.90 & 0.95 & 0.00 \\ 0.80 & 0.95 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.95 & 1.00 \end{pmatrix} \end{array}$$

$$FMT = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 1.00 & 0.80 & 0.85 & 0.30 \\ 0.80 & 1.00 & 0.00 & 0.00 \\ 0.85 & 0.00 & 1.00 & 0.00 \\ 0.30 & 0.00 & 0.00 & 1.00 \end{pmatrix} \end{array}$$

$$NEED = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{c} Q \end{array} \\ \begin{pmatrix} 0.00 \\ 0.90 \\ 0.00 \\ 0.00 \end{pmatrix} \end{array}$$

우선 퍼지 용어 매트릭스 FMT의 전이폐쇄 FMT_R_T를 [정의3.3]의 알고리즘을 이용하여 구하면 다음과 같이 구해진다.

$$FMT\_R_T = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 1.00 & 0.80 & 0.85 & 0.30 \\ 0.80 & 1.00 & 0.80 & 0.30 \\ 0.85 & 0.80 & 1.00 & 0.30 \\ 0.30 & 0.30 & 0.30 & 1.00 \end{pmatrix} \end{array}$$

구해진 FMT_R_T를 이용하여 식 (5)에 대입해 보면 전체적으로 다음과 같이 표현되고 결과 값이 구해진다.

$$REL = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 0.00 & 0.90 & 0.95 & 0.00 \\ 0.80 & 0.95 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.95 & 1.00 \end{pmatrix} \end{array} \circ$$

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 1.00 & 0.80 & 0.85 & 0.30 \\ 0.80 & 1.00 & 0.80 & 0.30 \\ 0.85 & 0.80 & 1.00 & 0.30 \\ 0.30 & 0.30 & 0.30 & 1.00 \end{pmatrix} \end{array} \circ$$

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{c} Q \end{array} \\ \begin{pmatrix} 0.00 \\ 0.90 \\ 0.00 \\ 0.00 \end{pmatrix} \end{array} = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{c} Q \end{array} \\ \begin{pmatrix} 0.90 \\ 0.90 \\ 0.80 \end{pmatrix} \end{array}$$

이상의 추론 결과에서 보는 것과 같이 사용자의 질의 Q(t_2, 0.9)에 대해 문서 d_1과 d_2에 대한 검색 결과는 논문[4-5]에서 제안된 식 (2)의 추론 방식의 검색 결과와 일치한다. 그러나 논문[4-5]에서 제안된 식 (2)의 추론 방식의 검색 결과에서는 합성 추론에 따른 퍼지 다언어 시소러스에 있는 묵시적인 링크(Implicit Link)를 무시하고 추론함으로써 문서 d_3이 사용자의 질의에 관련 없는 문서로 추론되어 검색되지 않았다.

이와는 다르게 본 논문에서 제안한 추론 방식에서는 위의 추론 결과에서 보는 것과 같이 문서 d_3도 사용자의 질의에 0.80 정도의 관련성을 가지고 있는 것으로 추론하게 된다. 이와 같이 본 논문에서 제안된 방법은 인간 전문가와 좀 더 유사하게 추론할 수 있다는 관점에서 논문[4-5]에서 제안된 추론 방법보다 좀 더 효과적이다.

IV. 결론 및 향후 연구 방향

단일언어 정보검색을 포함한 다언어 정보검색 시스템에서 검색 효과를 높이기 위한 하나의 방법으로 지식베이스인 시소러스를 이용하게 된다. 시소러스에 표현된 지식은 질의 평가시 평가 함수에 반영될 수 있기 때문에 시소러스의 구조는 평가 함수의 설계에 많은 영향을 준다. 이러한 관점에 따라 논문[4-5]에서는 퍼지 기법을 이용한 다언어 정보검색 시스템을 제안하였다. 논문[4-5]에서는 자연어에 기본적으로 존재하는 모호성을 해결하기 위하여 퍼지 다언어 시소러스를 이용함으로써 검색의 효과를 높인다.

그러나 논문[4-5]에서 제안된 추론 방식에서는 시소러스 내의 용어들 간의 묵시적인 링크를 무시함으로써 일부 검색되어야 하는 문서가 검색되지 않게 되는 문제점이 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 다언어 정보검색을 위한 퍼지 용어 매트릭스를 정의하고 정의된 퍼지 용어 매트릭스에 전이폐쇄 알고리즘을 적용하여 추론함으로써 인간 전문가에 좀 더 가까운 검색 결과를 얻을 수 있는 효과적인 다언어 문서 검색 방법을 제안하였다.

앞으로의 연구 방향은 제안된 방법에 따른 다언어 시소러스를 구축하고 이를 이용하여 웹 문서와 같은 실용적인 정보검색 분야에 적용해 보아야 할 것이다.

참 고 문 헌

[1] Radecki, T. Fuzzy set theoretical approach to document retrieval, Information Processing & Management, Vol. 15, pp. 247-259, 1979.

[2] Kar Wing Li and Rob Law, A novel english/chinese information retrieval approach in hotel website searching, Tourism Management, Vol. 28, pp. 777-787, 2006.

[3] Croft, W. B., Broglio, J. and Fujii, H., Applications of multilingual text retrieval, Proceedings of the 20th Annual Hawaii International Conference on System Sciences, pp. 98-107, 1996.

[4] Rowena Chau and Chung-Hsing Yeh, A fuzzy knowledge-based system for cross-lingual text retrieval, Computational Intelligence for Modelling, Control & Automation, M. Mohammadian(Ed.) IOS Press, pp. 488-494, 1999.

[5] Rowena Chau and Chung-Hsing Yeh, Crossing the language barrier using fuzzy logic, Springer-Verlag, pp. 768-773, 2005.

[6] Jacques savoy, Cross-language information retrieval; experiments based on CLEF 2000 corpora, Information Processing and Management, Vol 39, pp. 75-115, 2003.

[7] Timothy J. Ross, Fuzzy Logic with Engineering Applications, McGraw-Hill, Inc., 1995.

[8] M.R. Quillian, "Semantic memory," in Semantic information processing, M. Minsky ed., MIT Press, Cambridge Massachusetts, pp. 227-270, 1968.

[9] Kim. Y. W.; Kim, J. H. A model of knowledge based information retrieval with hierarchical concept graph, Journal of Documentation, 46(2), pp. 113-136, 1990.

[10] Lucarella, D.; Morara, R. FIRST: Fuzzy information retrieval system, Journal of Information Science, Vol. 17, pp. 81-91, 1991.

[11] Lee, J. H., Kim, M. H., & Lee, J. H. Ranking documents in thesaurus-based boolean retrieval systems, Inforamtion Processing & System. 30(1), pp. 79-91, 1994.

[12] Chen, S. M.; Wang, J. Y, Document retrieval using knowledge-based fuzzy information retrieval techniques, IEEE Transactions on systems, man, cybernetics, Vol. 25, No. 5, pp. 793-803, 1995.

[13] Kandel, A. Fuzzy mathematical techniques with applications. CA: Addison-Wesley, 1986.

저 자 소 개

최명복(Myeong-Bok Choi)

- 1992년 : 호서대학교 전자계산학과(학사)
- 1994년 : 아주대학교 컴퓨터공학과(석사)
- 2001년 : 아주대학교 컴퓨터공학과(박사)
- 현재 : 강릉대학교 컴퓨터공학부(부교수)
- 2003. 9～현재 한국정보과학회 전산교육연구회 운영위원
- 2003. 9～현재 한국정보과학회 전문대학 학회지·논문지 편집위원
- 2004. 1～현재 한국컴퓨터산업교육학회 학회지 편집위원
- 2004. 1～현재 한국인터넷방송/TV학회 협동이사

<주관심분야 : 지능형 정보검색, 퍼지응용, 지식표현, 신경망, 지능형 교통제어, 소프트웨어 공학, 임베디드 응용 시스템>