요구 사항 문장 범주화를 이용한 웹 기반의 요구 사항 추출 지원 시스템

(Web-based Requirements Elicitation Supporting System using Requirements Sentences Categorization)

고 영 중 † 강 기 선 † 김 재 선 † 박 수 용 †† 서 정 연 †††

(Youngjoong Ko) (Kisun Kang) (Jaeseon Kim) (Sooyong Park) (Jungyun Seo)

요 약 시스템이 사용되는 분야가 점점 복잡해지고 대형화됨에 따라 시스템 개발에 있어 사용자 요구 사항의 올바른 분석과 서술이 중요하게 인식되고 있으며, 인터넷(internet)의 발전으로 분산 환경에서의 요구 사항 추출 및 분석의 필요성이 대두되고 있다. 본 논문에서는 자연어로 표현되는 요구 사항 문장을 유사도 측정 기법을 이용하여 주제별로 범주화(categorization)함으로써 분산 환경에서 수집된 요구 사항 문장을 분석하기 위한 기초를 제공할 수 있는 요구 사항 추출 지원 시스템을 제안한다. 제안된 시스템은 단어간, 문장간의 유사도 측정 기법을 이용하여 수집된 요구 사항 문장들을 주제별로 자동으로 분류함으로써 요구 사항 분석 시 초기 작업의 어려움을 줄이고 신속하고 정확하게 분석 작업을 수행하도록 지원할 것이다. 본 논문에서는 단어간, 문장간 유사도 측정 기법을 이용한 범주화 기법의 효율성을 실험을 통해 검증하였으며 구현된 시스템을 통해 추출, 처리되는 과정을 보여주고 있다.

Abstract As a software becomes more complicated and large-scaled, it is very important for a software engineer to analyze user's requirements precisely and apply them effectively in the development stage. Due to the growth of the internet, the necessity of requirements elicitation and analysis in distributed environments has also become larger. This paper proposes a requirements elicitation supporting system that offer the basis for effectively analyzing requirements collected in distributed environments. The proposed system automatically categorizes collected requirements sentences into selected subject fields by measuring their similarity using a similarity measurement technique. Therefore, it reduces the difficulties in the initial stage of requirements analysis and it supports rapid and correct requirements analysis. This paper verifies the efficiency of the proposed system in similarity measurement techniques through experiments, and presents a process for requirements specifications elicitation using the embodied system

1. 서 론

시스템이 사용되는 분야가 점점 복잡해지고 대형화됨에 따라 시스템 개발에 있어서 사용자 요구 사항의 올바른 분석과 서술이 점차 중요한 분야로 부각되기 시작하였다. 이에 따라 90년대에 들어오면서 요구 사항에 관계되는 모든 활동과 원칙들을 요구 공학(Requirements Engineering)이라 하여 많은 연구가 진행되고 있다[1][2]. 이렇게 복잡하고 대형화되는 시스템 개발에 있어서는 다양한 부류의 사용자들이 존재하므로 이들의 효과적인 참여를 통한 요구 사항의 추출 및 분석이 시스템 개발의 중요한 성공 요소가 된다. 그러므로, 요구 사항의 효과적인 추출과 추출된 요구 사항의 효과적인 분석

· 본 연구는 교육부의 BK21 사업 핵심연구과제를 통한 지원으로 이루어 진 것입니다.

† 비 회 원 : 서강대학교 컴퓨터학과
kyj@nlpzodiac.sogang.ac.kr
kadin@selab.sogang.ac.kr
jskim@selab.sogang.ac.kr
†† 정 회 원 : 서강대학교 컴퓨터학과 교수
sypark@ccs.sogang.ac.kr
††† 종신회원 : 서강대학교 컴퓨터학과 교수
scojy@ccs.sogang.ac.kr

논문접수 : 1999년 7월 30일
심사완료 : 2000년 2월 9일

을 위해서 대화 분석(dialogue analysis)이나 요구 사항 범주화(categorization) 혹은 클러스터링(clustering)등의 여러 가지 자연어 처리 기법들(natural language processing techniques)이 응용되어 왔다[3][4].

다양한 부류의 사용자들로부터의 효과적인 요구사항 추출은 공동의 사용 환경이 필요한데 이것의 해결 방안 중의 하나가 인터넷이라 할 수 있다. 인터넷(internet)의 발전과 보급으로 인해 웹(web)은 친숙한 인터페이스가 되어가고 있으며 이는 서로 상이한 특성과 목적을 가진 여러 그룹들이 분산 환경에서 다른 시간에 컴퓨터를 이용한 공동작업으로 소프트웨어 개발에 협력하고 참여할 수 있는 새로운 가능성을 만들어 주고 있다. 작업의 성격이 변화되고 시스템이 복잡해짐에 따라 이러한 분산 환경에서의 공동작업의 필요성이 증대되었으며 또한 중요하게 인식되어 왔다[5]. 이와 같이 웹을 이용한 분산 환경에서의 작업의 필요성은 고객, 시스템 사용자, 그리고 시스템 개발에 관련된 사람들이 서로 의견을 교환함으로써 실제로 개발하고자 하는 시스템에 대한 요구들을 찾아내고자 하는 요구 공학의 첫 번째 공정인 요구 사항 추출 단계에서의 웹 사용의 요구를 뒷받침하고 있다[5][6]. 요구 사항 추출 단계에서 웹의 사용은 요구 사항 추출 작업의 시간적, 공간적 제약을 극복하게 해주지만 전통적인 요구 사항 추출 기법들인 scenario, questionary, interview, conversation 등[7]과는 달리 요구 사항들이 분산되어 수집됨으로써 분석단계 초기에 전통적인 기법보다는 더 많은 어려움이 있게 된다. 시스템 개발 시 다양한 부류의 사용자들이 웹을 통하여 그들의 요구 사항을 제공할 경우 이러한 요구 사항들은 그들의 시스템에 대한 한가지 관점만을 표현하므로 이러한 단편적 요구 사항들이 한꺼번에 웹을 통하여 시스템 개발자에게 전달이 될 경우 이러한 것들의 효과적인 분석이 매우 어려운 상황이 된다. 예를 들어 시스템 개발의 경우 40명의 사용자로부터 각 사용자 당 50개의 요구 사항이 제안되어 총 2000개의 요구 사항이 수집되었다면 이들의 관점이 제각기 다르고 내용 자체가 분산되어 있어 이들의 일관적인 분석이 매우 어려운 상황이 될 것이다. 그러므로, 요구 사항 분석 단계의 준비 단계에서 수집된 요구 사항 문장들을 주제별로 자동으로 범주화한다면 분석 초기 단계의 어려움을 줄이고 분석을 위한 좋은 기초를 제공할 수 있을 것이다[8].

본 논문에서는 단어간, 문장간의 유사도(similarity)기법을 이용하여 요구 사항 문장들을 범주화함으로써 웹을 기반으로 한 요구 사항의 추출과 분석을 지원하는 자동화된 도구의 개발을 제안한다. 제안된 도구는 [그림 1]과 같이 요구 공학의 추출과 분석 공정 단계에서 다양한 작업 그룹 참여의 시간적, 공간적 제약을 극복하고 요구 사항 분석을 위한 기초를 제공할 것이다.

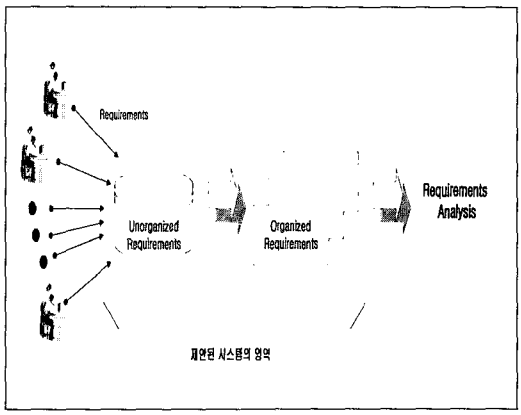

그림 1 제안된 시스템을 이용한 요구 사항 분석 과정

본 논문의 구성은 다음과 같다. 먼저 2장에서 관련된 연구들을 설명한다. 그리고 3장에서는 시스템의 구조를 간략하게 보이고, 문장들의 범주화에 사용된 문장간의 유사도 측정 기법에 대해 자세하게 설명한다. 4장에서는 실험을 통하여 요구 사항 문장의 범주화의 정확도와 효율성을 평가한다. 5장에서는 앞장에서 제시한 문장간 유사도 측정 기법을 이용한 요구 사항 문장 범주화 시스템의 구현 및 결과를 분석한다. 마지막으로 6장에서 결론을 내리고, 향후 연구 과제를 제시한다.

2. 관련 연구

요구 사항 문서를 자동으로 분류하는 기존의 연구들은 주로 문서에서 추출된 특정한 키워드들과 유의어 사전(thesaurus)을 이용하거나 단어의 반복(reiteration)과 공기 정보(collocation) 같은 언어 현상을 이용한 간단한 유사도 측정 방법을 이용하였다[9]-[12]. Palmar는 [11]에서 2층 구조의 TTC(two-tired clustering)알고리즘을 이용하여 요구 사항 문서를 색인하고 클러스터링 하는 방법을 제안하였다. TTC알고리즘은 먼저 각 요구 사항 문서에 속해 있는 동사들을 키워드로 사용하고 동사 유의어 사전을 사용하여 문서를 기능별로 분류하고, 이렇게 기능별로 분류된 문서들 사이에 코사인(cosine) 유사도를 측정하여 재분류하는 방법이다. 그리고, Yaung은 [12]에서 그래프 모델을 기반으로 하여 결합(cohesion) 개념을 사용해서 높은 결합 상태를 가지는 요구 사항들은 같은 클러스터 안에 위치하고 낮은 결합

상태를 가지는 요구 사항들은 서로 다른 클러스터에 위치하게 함으로써 요구 사항 클러스터링 분석을 제안하였다. 그러나, 이러한 방법들은 요구 사항 문서들을 클러스터링 하는 기법으로서 요구 사항 문서에 비해 적은 의미 정보를 가지고 있는 문장 단위의 요구 사항들의 주제별 범주화에는 효율적인 방법이 되지 못한다. 또한, 유의어 사전을 이용한 방법들은 문서간 유사도 측정의 정확율 면에서 개선된 성능을 보이고는 있으나 특정 영역마다 이러한 사전을 구축하고 관리하는 일은 결코 쉬운 작업이 아니다.

Lecoeuche는 [3]에서 자연어 대화(natural language dialogue)를 사용해서 요구 사항을 추출하기 위한 이론을 제시하고 있으며 Parmer은 [4]에서 MDGSS (Multigroup Decision Support System)을 구현하고 이를 이용해서 다양한 부류의 사용자들로부터의 효과적인 요구 사항 추출과 분석을 수행하고자 하였다. 하지만 이러한 공동 작업을 위한 시스템(collaborative system)이 인터넷상에서 구현된다면 서로 상이한 특징과 목적을 가진 여러 그룹들(nonhomogeneous group)이 분산된 장소에서 다른 시간(distributed and asynchronous)에도 인터넷을 이용하여 보다 쉽고 편리하게 시스템 개발에 협력하고 참여할 수 있을 것이다[5].

그러므로, 본 논문에서는 단어간, 문장간의 유사도 측정 기법을 이용하여 유의어 사전을 사용하지 않고 요구 사항 문장의 범주화를 수행하는 웹 기반의 요구 사항 추출 지원 도구를 제안한다.

3. 요구 사항 문장 범주화 기법

3.1 범주화 시스템 구조도

본 논문에서 제안한 범주화 시스템의 구조도는 [그림 2]와 같다. 시스템은 크게 주제별, 키워드별 예제 추출부와 문장간 유사도 측정부로 나눌 수 있다. 여기서 주제어 예제란 수집된 요구 사항 문장들 중에서 주제어를 직접 포함하고 있는 문장을 말하고 키워드(keyword) 예제는 요구사항 문장들 중에서 키워드를 직접 포함하고 있는 문장을 말하는 용어이다. 본 논문에서는 이들 주제어 예제와 키워드 예제들이 각 주제의 의미를 가장 잘 반영하고 있는 문장으로 간주한다.

주제별, 키워드별 예제 추출부는 먼저 형태소 분석 및 태깅 작업을 통해 수집된 요구 사항 문장에서 각각의 주제어나 키워드를 포함하고 있는 문장을 추출하는 부분이고 문장간 유사도 측정부는 주제별, 키워드별 예제 추출부에서 추출되어 주제별로 분류된 문장들과 미분류 요구 사항 문장들과의 유사도 측정을 통해 미분류 요구 사항 문장을 가장 유사도가 높은 문장이 있는 주제 범주로 할당하는 부분이다.

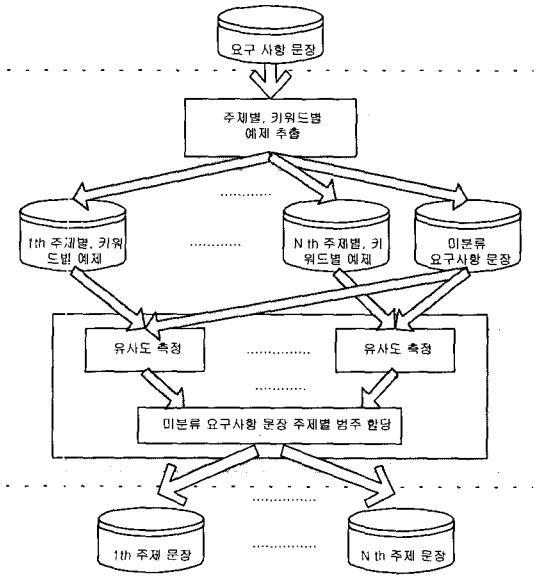

그림 2 범주화 시스템 구조도

3.2 주제별, 키워드별 예제 추출

수집된 요구 사항 문장들을 각 주제별 범주화를 하기 위해서는 먼저 각 범주들의 특징들을 가장 잘 포함하는 문장들을 추출해야 한다. 본 논문에서는 사용자가 입력한 범주별 주제어와 키워드를 사용하여 문장들을 추출하였으며 이들이 주제어 예제와 키워드 예제이다. 키워드는 주제어와 유사도가 높은 단어들로 구성되는데 기존의 연구에서는 유의어 사전[11]이나 전자 사전(machine-readable dictionary)[13]을 이용하였다. 그러나, 요구 사항 문장들에서는 영역에 따라 단어들이 특징지어 지는 경우가 많아 일반적인 특징을 표현하고 있는 유의어 사전이나 전자 사전을 사용해서는 높은 성능을 기대하기 어렵다. 또한, 영역마다 새로 유의어 사전을 구성하거나 전자 사전을 구성한다는 것도 어려운 작업임이 분명하며 요구 공학 분야에서 일반적으로 사용할 수 있는 유의어 사전이나 전자 사전의 구축에 관한 연구가 미흡한 실정이다.

3.2.1 내용어 추출

주제별, 키워드별 예제를 추출하기 위해서 그 문장의 내용이나 특징을 잘 내포할 수 있는 단어를 추출해야 하는데 이러한 단어를 내용어(content word : open-class word)라고 한다.

한국어에서는 동사인 경우에 동작성 명사에 '-하다', '-되다' 등의 동사 파생 접미사가 붙어서 동사가 되는 경우가 많고 형용사의 경우에도 상태성 명사와 더불어 '-하' 등의 형용사 파생 접미사가 붙어서 형용사가 되는 경우가 많으므로 이와 같은 종류의 동사와 형용사는 한국어 형태소 분석기에서는 명사로 추출된다. 그리고, 이 밖의 동사나 형용사는 '하다', '되다', '있다', '없다' 등 불용어에 해당하는 의미를 갖는 경우가 많다. 또한, 본 논문의 연구 영역인 요구 사항 문장에서는 동사와 형용사가 내용어로서 의미를 갖는 경우가 더욱 적어지는데 실제로 본 논문에서 실험을 위해 사용한 요구 사항 문장(총 180문장)들을 분석해 본 결과 총 449개의 동사와 형용사중에서 동작성명사나 상태성명사로 추출되는 경우가 267개(60%)이고 그 외의 182개의 동사, 형용사중에 86개(47%)가 '하다', '되다', '있다', '없다'등의 불용어에 해당하는 의미를 갖는 경우이다. 그리고, 그 외의 다른 동사나 형용사도 요구 사항 문장 특성상 문장의 내용이나 특징을 잘 나타내지 못하는 단어들('받다', '어렵다' 등)로 구성된다. 그러므로 본 논문에서는 한국어의 특성과 연구 영역에 맞게 명사만을 이용해서 내용어를 추출하였다.

이렇게 추출된 내용어에는 여러 문장에서 공통적으로 많이 발생하기 때문에 별다른 정보를 주지 못하는 단어들이 있는데 '처리', '부분' 등이 이에 해당한다. 본 논문에서는 이를 처리하기 위해 불용어 사전을 두어 내용어 추출 시 불용어에 해당하는 단어를 제거하였다.

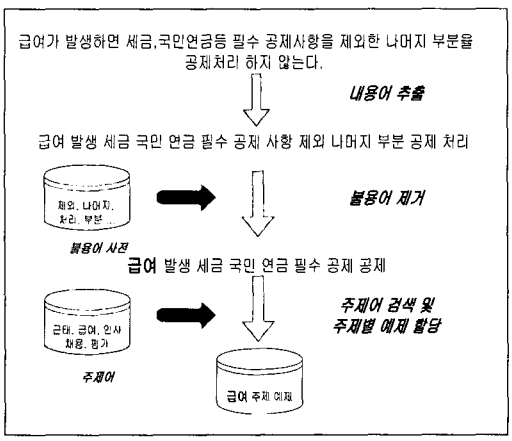

그림 3 주제어 예제 추출의 예

3.2.2 예제 문장 추출

추출된 요구 사항 문장의 내용어 중에 먼저 주제어에 해당하는 내용어를 가진 주제어 예제를 추출하고 다음으로 키워드에 해당하는 내용어를 가진 키워드 예제를 추출한다. 이때 두가지 이상의 주제나 키워드에 해당하는 내용어를 가진 문장은 예제 문장 추출에서 제외시킨다. [그림3]는 주제어 예제 추출의 예이다.

3.3 단어간, 문장간 유사도 측정

3.2절에서 추출된 주제별, 키워드별 예제들과 미분류 요구 사항 문장들과의 문장간의 유사도 측정을 통해 미분류 요구 사항 문장들을 각 주제 범주로 할당하게 된다.

유사한 단어는 유사한 문맥(context)에 위치하는 경향이 있으므로 이를 이용해 문맥 정보를 반영하여 문장간 유사도를 측정하였다[9][13]. 본 논문에서 단어와 문장은 상호 보충적인 역할을 수행한다. 문장은 포함하고 있는 단어들에 의해 표현되고 단어는 그 단어를 포함하고 있는 문장들에 의해 표현된다. 즉, 문장은 유사한 단어들을 많이 포함할수록 유사한 문장이고 단어는 유사한 문장에서 많이 사용될수록 유사한 단어이다. 이 정의는 순환적이며 이를 반영하기 위해 [그림4]과 같이 두 개의 행렬(matrix)을 이용해서 반복 계산한다.

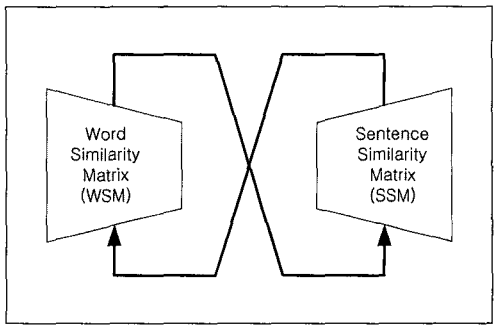

그림 4 단어, 문장 유사도 측정의 반복 계산

[그림4]의 *WSM*(word similarity matrix)의 행(row)과 열(column)은 주제별로 주제어 예제, 키워드 예제들과 미분류 요구 사항 문장들에 포함되어 있는 모든 내용어(content word)들로 구성되며 행렬의 각 요소(cell)는 단어사이의 문맥적 유사도를 나타내는 0에서 1 사이의 값을 가진다. *SSM*(sentence similarity matrix)은 행에 미분류 요구 사항 문장들이 위치하고 열에 주제어, 키워드 예제들이 위치함으로써 이들 문장간의 유사도 값을 나타내게 한다.

이들 단어간, 문장간 유사도를 측정하기 위해서는 먼저 *WSM*을 단위 행렬(identity matrix)로 초기화한다.

즉, 각 단어는 같은 단어의 유사도 값은 1로 하고 다른 단어와의 유사도 값은 0으로 한다. 그리고, 유사도 값의 변화가 충분히 작아질 때까지 다음과 같은 과정을 반복 수행한다.

1. WSM_n을 사용해서 SSM_n을 갱신(update)한다.
2. SSM_n을 사용해서 WSM_n을 갱신(update)한다.

3.3.1 친밀도 계산식

단어와 문장의 유사도 측정 계산을 단순화하기 위해서 단어와 문장 사이의 관계를 정의하는데 이를 친밀도(affinity)라 한다. 단어(W)는 모든 문장과 친밀도를 가지는데 이는 단어(W)와 문장을 구성하고 있는 단어들과의 문맥적 관계를 나타낸다. 만약 단어(W)가 문장에 속해 있다면 친밀도는 1이고 문장(S)의 단어들과 전혀 관계가 없다면 친밀도는 0에 가깝고 단어(W)가 문장(S)의 단어들과 문맥적으로 유사하다면 0에서 1의 값을 가지게 된다. 비슷한 방법으로 문장(S)도 모든 단어와 친밀도를 가질 수 있는데 이는 문장(S)와 단어(W)를 포함하고 있는 문장들과의 유사도를 반영한다.

친밀도 계산식은 다음과 같다[13]. 여기서 단어(W)가 문장(S)에 속해 있을 때 $W \in S$로 표시한다.

$$aff_n(W, S) = \max_{W_i \in S} sim_n(W, W_i) \tag{1}$$

$$aff_n(S, W) = \max_{S_i \ni W} sim_n(S, S_i) \tag{2}$$

n은 반복 횟수를 나타내며 유사도 값은 WSM_n과 SSM_n에 의해 정의된다. 이 식에 의해 모든 단어는 모든 문장과 친밀도에 의해 표현될 수 있고 문장은 포함하고 있는 단어들의 친밀도를 나타내는 벡터(vector)들로 표현된다. 그러나, 일반적으로 친밀도식은 [식3]과 같이 비대칭적(asymmetric)이다.

$$aff_n(W, S) \neq aff_n(S, W) \tag{3}$$

3.3.2 유사도 계산식

단어 W_1과 W_2의 유사도는 W_1과 W_2를 포함하고 있는 문장들의 평균 친밀도에 의해 정의되고 문장 S_1와 S_2의 유사도는 S_1와 S_2를 구성하는 단어들의 친밀도의 가중치(weight) 평균으로 정의된다. 유사도 계산식은 다음과 같다[13].

$$sim_{n+1}(S_1, S_2) = \sum_{W \in S_1} weight(W, S_1) \cdot aff_n(W, S_2) \tag{4}$$

$$\text{if } W_1 = W_2 \quad sim_{n+1}(W_1, W_2) = 1$$
$$\text{else} \quad sim_{n+1}(W_1, W_2) = \sum_{S \ni W_i} weight(S, W_1) \cdot aff_n(S, W_2) \tag{5}$$

여기서 [식4]의 가중치는 3.3.3절에서 기술한 방식을 따랐으며 [식5]의 가중치는 합이 1이 되도록 단어 W_1를 포함하고 있는 문장 개수의 역수를 사용하였다. 이 식에서 계산된 유사도 값은 WSM과 SSM의 각 요소에 대응되는 값이다.

3.3.3 단어 가중치

[식4]에서 사용한 단어의 가중치는 다음과 같은 3가지 요소에 의해 결정된다. 가중치는 자질(feature)의 수를 줄이고 유사도 값에 기여하지 않는 단어를 제외시키는데 사용되며 반복계산 중에는 변경되지 않는다.

1. 요구 사항 문장에서의 출현 빈도 : 전체 요구 사항 문장에서 자주 등장하는 단어는 문장의 유사도나 의미의 정보량을 적게 지니게 된다. 예를 들어 '관리' 같은 단어는 어떤 요구 사항 문장에도 자주 나오게 된다. 이를 반영하기 위한 식은 다음 [식6]과 같다[13].

$$max\left\{0, 1 - \frac{freq(W)}{\max 5\_freq(x)}\right\} \tag{6}$$

$\max 5_freq(x)$는 전체 요구 사항 문장에서 가장 출현 빈도가 높은 5개의 출현 빈도를 합한 값이다.

2. 로그 확률 요소(log-likelihood factor) : 주제어의 의미를 잘 나타내는 단어는 전체 요구 사항 문장에서의 출현 확률보다 주제별, 키워드별 예제에서 더 출현 확률이 높다. 이는 [식7]에 의해 계산된다[13].

$$log \frac{\Pr(W_i|W)}{\Pr(W_i)} \tag{7}$$

$\Pr(W_i)$는 전체 요구 사항 문장에서의 단어 W_i의 출현 확률이며 $\Pr(W_i|W)$는 주제별, 키워드별 예제에서의 W_i의 출현 확률이다. 주제별, 키워드별 예제에서의 출현 빈도가 낮은 경우를 반영하기 위해 [식7]에서 계산된 값에 $\min\left\{1, \frac{count(W_i)}{3}\right\}$을 곱하였다. 여기서 $count(W_i)$는 예제 문장에서의 W_i의 출현 빈도이다. 본 논문에서는 예제 문장에서 출현하지 않은 단어들에 값을 할당하기 위해 이들의 가중치 값은 1로 하고 예제 문장에 출현한 단어는 [식7]에서 계산된 값에 1을 더해서 사용하였다.

3. 품사(part of speech) : 추출된 내용어의 품사에

따라 다른 가중치를 주었는데 본 논문에서는 고유명사, 외국어, 일반명사인 경우에는 가중치를 1로 하였으며 동작성 명사, 상태성 명사인 경우에는 0.6으로 하였다.

위의 3가지 가중치 값의 곱을 어떤 단어의 가중치로 사용하였으며 문장안에서 각 단어는 [식8]의 식으로 정규화(normalization)하였다[13].

$$weight = \frac{factor(W_i, S)}{\sum_{W_i \in S} factor(W_i, S)} \quad (8)$$

$factor(W_i, S)$는 정규화 전의 가중치 값이다.

4. 실험 및 평가

4.1 실험 방법 및 데이타 구성

본 논문에서 제안하는 문장간 유사도 측정 기법을 이용한 요구 사항 문장 범주화 기법은 분산 환경에서 입력되는 요구 사항 문서들을 효율적으로 분류하고 분석하는 작업에 사용될 수 있다. 제안된 문장간 유사도 측정 기법의 효율성을 평가하기 위하여 다음과 같이 실험하였다. 먼저 수집된 요구 사항 문장을 5개의 주제('근태', '급여', '인사', '채용', '평가')별로 수작업으로 분류하여 정답으로 사용하였으며 실험에서 사용한 데이타는 총 180문장(1829어절, 10.16어절/문장)이다. [표1]은 실험에서 사용한 주제어와 키워드들이다.

표 1 실험에서 사용한 주제어, 키워드

주제어	키워드
근태	출퇴근, 야근, 휴가
급여	보수, 수당, 호봉, 연봉, 월봉
인사	승진, 징계, 직급, 직위, 근속
채용	전형, 면접, 합격, 시험
평가	점수, 실적, 능력, 조정

4.2 실험 결과

본 논문에서는 성능을 평가하기 위하여 정보 검색 분야(information retrieval)에서 일반적으로 사용되는 정확율(precision)과 재현율(recall)을 사용하였다. 그 식은 다음과 같다[14].

$$precision = \frac{category\ sentences\ found\ and\ correct}{total\ category\ sentences\ found} \quad (9)$$

$$recall = \frac{category\ sentences\ found\ and\ correct}{total\ category\ sentences\ correct} \quad (10)$$

[표2]은 주제별 정확율과 재현율에 대한 표이고 [그림5]는 이를 비교하기 위해 도식화한 것이다.

표 2 범주화 기법의 정확율과 재현율 비교

	근태	급여	인사	채용	평가	total
정확율	96.88 (31/32)	86.44 (51/59)	74.00 (37/50)	91.67 (11/12)	85.19 (23/27)	85.00 (153/180)
재현율	79.49 (31/39)	87.93 (51/58)	84.09 (37/44)	100.00 (11/11)	82.14 (23/28)	

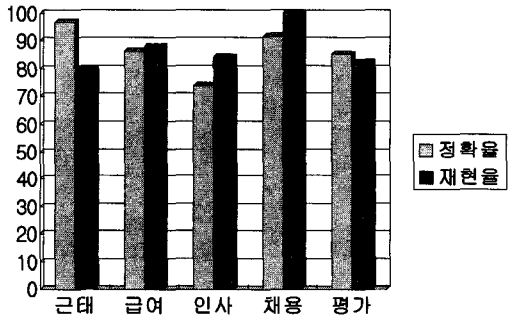

그림 5 범주화 기법의 정확률과 재현률 비교

[그림5]를 보면 정확율과 재현율이 '근태'의 주제어인 경우를 제외하고는 큰 차이를 보이고 있지 않다. '근태'의 경우에 정확율에 비해 재현율이 낮은 이유는 '급여'와 '인사'등의 다른 주제어와 의미 자체가 비슷하기 때문에 요구 사항 문장들이 다른 주제 범주로 많이 분류되었기 때문으로 분석된다.

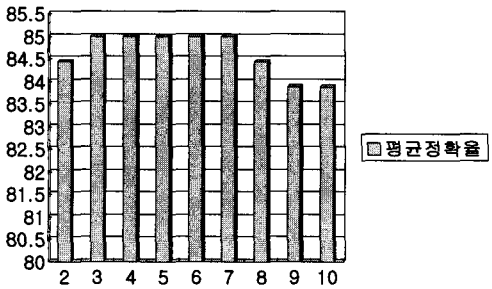

그림 6 반복 횟수에 따른 평균 정확율의 비교

유사도 측정 계산의 반복 횟수를 결정하기 위해 [그림6]와 같이 반복 횟수를 달리해서 평균 정확율을 비교해 보았다.

반복 횟수는 [그림6]에서와 같이 3~7회에 가장 높은 평균 정확율을 보였으며 반복 횟수의 수가 많아질수록 수행 속도가 느려지므로 본 실험에서는 반복 횟수를 3회로 결정하여 실험하였다.

5. 구현 및 평가

본 절에서는 제안된 기법들을 실제 시스템에 적용한 구현의 모습과 평가들을 기술한다. 시스템은 Solaris상에서 C와 CGI를 사용해서 구현했으며 제안된 웹기반 요구 분석 시스템의 전체 구성도는 [그림7]과 같다.

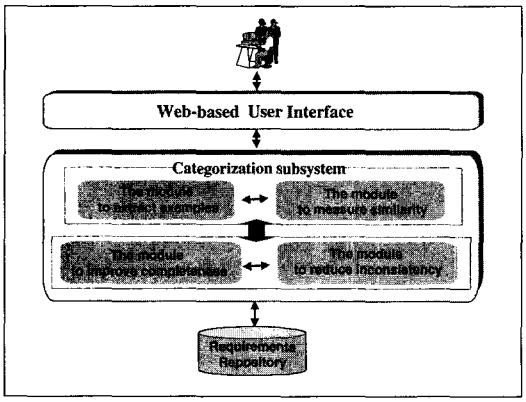

그림 7 웹기반 요구 분석 시스템의 전체 구성도

자료를 저장하고 주요 기능을 수행하는 것은 서버에서 수행되며 클라이언트는 서버로부터 자료를 받아서 사용자에게 보여주고 사용자가 입력한 자료를 서버로 보내주는 역할만을 수행한다. 그러므로, 등급에 따라 권한이 제한되어 있는 사용자들이 한 저장소를 중심으로 소프트웨어 개발에 참여할 수 있기 때문에 일관성이 보장된다.

본 도구의 사용자는 프로젝트 관리자(project manager), 시스템 엔지니어(system engineer), 그리고 일반 사용자들로 나누어지는데 이들은 부여받은 권한 등급에 따라 로그인하고 자신에게 허용된 기능을 실행할 수 있다. 프로젝트 관리자가 프로젝트를 생성하고, 관련자들을 사용자로 등록한 후 시스템 엔지니어와 일반 사용자들간의 의견 교환을 통해 일반 사용자가 자신들의 요구 사항을 입력, 수정, 삭제한다. 그리고, 시스템 엔지니어가 본 논문에서 제안한 범주화 기법을 이용해서 수집된 요구 사항들을 분석을 위한 구조로 분류해 낸다. 최종적으로 [9]에서 제안한 유사도 검사와 모호성, 중복성, 불일치성을 검사하는 기법 등을 사용한 지원 도구를 사용하여 요구 사항들을 재구성하고 분석하여 최종의 요구 사항 문서를 완성하게 된다. [그림8]은 제안된 시스템의 사용자 요구 사항 입력을 위한 초기 화면이다.

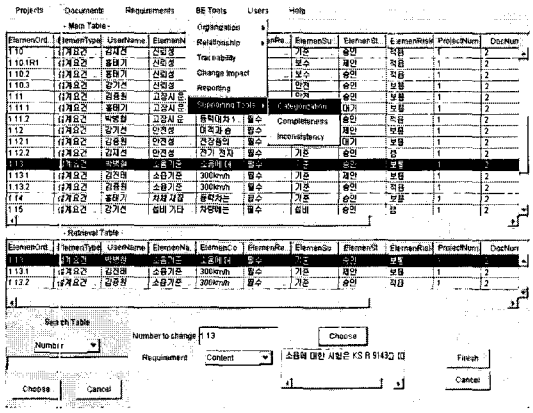

그림 8 사용자 요구 사항 입력을 위한 초기 화면

문장 범주화 서브 시스템의 수행 단계는 다음 3가지 단계로 나누어진다. 시스템 엔지니어는 먼저 'Input Subject'를 선택해서 주제어와 키워드를 입력하여 범주 영역을 결정하고 다음으로 'Processing'을 선택해서 요구 사항 문장 범주화 기능을 실행한다. 마지막으로 'Result'를 통해 실행 결과를 확인한다.

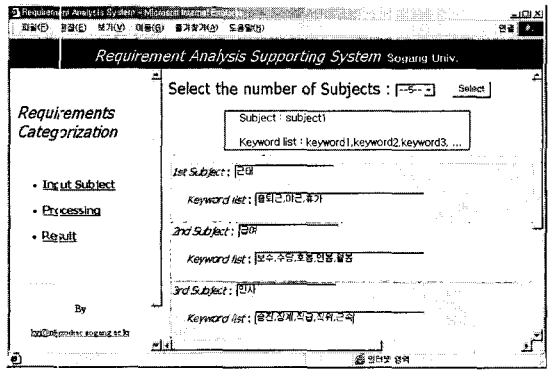

그림 9 주제어, 키워드 입력 사용자 인터페이스

[그림9]는 본 시스템의 주제어, 키워드 입력 사용자

인터페이스를 보이고 있다. 시스템 엔지니어는 먼저 주제 범주의 개수를 입력하고 각 범주의 주제어와 키워드를 입력한다.

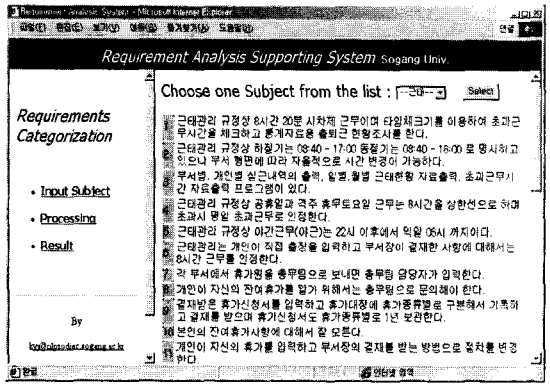

그림 10 요구 사항 문장 범주화 실행 결과

[그림10]은 입력된 주제별로 범주화 실행 결과의 요구 사항 문장을 보여 주고 있다. 시스템 엔지니어는 범주화 결과를 분석함으로써 자신이 선택한 각 범주의 주제어와 키워드를 수정해서 시스템 엔지니어가 원하는 대로 좀 더 나은 범주화 결과를 만들어 낼 수 있을 것이다.

여기서 문장 범주화 시스템을 통해 생성된 각 범주별 요구 사항들은 요구 사항 분석의 초기 단계의 어려움을 줄이고 분석을 위한 좋은 기초 자료로 사용될 수 있을 것이다.

6. 결론 및 향후과제

본 논문에서는 유사도 측정 기법을 이용하여 요구 사항 문장을 범주화함으로써 분산 환경에서 수집된 요구 사항 문장을 분석하기 위한 기초를 제공할 수 있는 요구 사항 추출 지원 시스템을 제안하였다. 제안된 시스템은 언어 분석의 비교적 하위 단계인 형태소 분석만을 이용하여 유사도를 측정하기 때문에 비교적 단순하고 쉽게 구현될 수 있다는 장점이 있다. 제안된 시스템을 이용한다면 분석 초기단계에서 요구 사항 문장들을 주제별로 자동으로 분류할 수 있기 때문에 요구 사항 문장들을 수작업으로 분류하는 것보다 작업량을 줄일 수 있으며 이로 인해 보다 신속하고 정확하게 분석 작업을 수행할 수 있을 것이다. 특히, 웹의 발달로 분산 환경에서의 작업이 보편화 되어가는 요즘에는 제안된 시스템의 효율성이 더욱 중요시될 것이다.

제안된 시스템을 통하여 분산된 관점과 내용의 요구 사항들을 범주화함으로써 요구 사항 분석을 매우 용이하게 지원 할 수 있다. 중요한 것은 범주화의 구조를 설정하는 것이며 이는 개발하고자 하는 영역 분석에 의하여 설정된다. 일단 범주 구조가 설정되면 이 구조에 의하여 요구 사항들을 분류하고 요구 사항 분석자들은 분류된 요구 사항을 사용해서 분석의 효율성을 높일 수 있게 된다.

향후 연구로는 주어진 분산된 요구 사항들로부터 자동으로 공통된 영역의 범주화 구조를 추출하는 것인데 이는 이미 개발된 범주화 구조의 재사용을 통하여 범주화 구조를 확장해가는 방법에 의하여 개발하고자 한다. 그리고, 본 논문에서 제안한 범주화 기법에서는 입력되는 키워드의 선택이 전체 시스템의 성능에 큰 영향을 미치게 되는데 키워드 선택에 도움을 줄 수 있는 방법의 개발도 필요할 것으로 보인다. 또한, 제안된 시스템의 실험을 위한 충분한 실험 데이타(요구사항 문장)를 확보하기가 어려워 실험이 충분히 되지 못하였다. 좀 더 대용량의 실험 데이타를 확보할 수 있다면 제안된 시스템의 성능을 좀 더 정확히 평가할 수 있었을 것이며 더욱 발전시킬 수 있을 것이다.

참 고 문 헌

[1] Ian Sommerville and Pete Sawyer, *Requirements Engineering,* Wiley,1997

[2] Richard H.Thayer and Merlin Dorfman, *Software Requirements Engineering,* 2nd Edition. IEEE Computer Society Press. 1997

[3] Renaud Lecoeuche, Chris Mellish, and Dave Robertson, "A Framework for Requirements Elicitation through Mixed-Initiative Dialogue," International Conference on Requirements Engineering(ICRE'98), 1998.

[4] James D. Palmer, N. Ann Fields, and Peggy Lane Brouse, "Multigroup Decision-Supprot Systems in CSCW," *Computer*, vol. 27, 1994. pp.67-72.

[5] James D. Palmer, N. Ann Fields, "Computer Supported Cooperative Work," *Computer*, vol.36, 1994

[6] 이원우, 황만수, 박수용, "웹을 이용한 요구사항 관리 모델의 구축", 정보과학회 가을 학술 발표 논문집, 1998

[7] Joseph A. Goguen and Charlotte Linde "Techniques for Requirements Elicitation," *Proceedings of the International Symposium on Requirements Engineering*, 1993

[8] Park, S. and Palmer, J., "Automated Support to System Modeling from Informal Software Requirements," *Proceedings of the 6th International*

conference on Software Engineering and knowledge Engineering, June 1994.
[9] 김학수, 고영중, 박수용, 서정연 "문서간 유사도 측정을 통한 효율적인 사용자 요구 분석", HCI'99 학술대회 논문집, pp.73-79, 1999
[10] Maarek, Y., Berry, D. and Kaiser, G., "An Information Retrieval Approach For Automatically Construction Software Libraries", *IEEE Transaction On Software Engineering,* Vol. 17, No. 8, pp.800-813, August 1991.
[11] Palmer, J. and Liang, Y., "Indexing and clustering of software requirements specifications," *Information and decision Technolgies,* Vol 18, pp. 283-299, 1992.
[12] Alan T. Yaung, "Design and Implemetation of a Requirements Clustering Analyzer for Software System Decomposition," *ACM* 1992
[13] Yeal Karov and Shimon Edelman, "Similarity-based Word Sense Disambiguation," *Computational Linguistics,* Vol 24, No 1, pp41-60, March 1998.
[14] Yang,Y., "An evaluation of statistical approaches to text categorization," *Information Retrieval Journal,* May, 1999.

고 영 중
1996년 서강대학교 수학과 학사. 1996년 ~ 1997년 LG-EDS system 근무. 2000년 서강대학교 컴퓨터학과 석사. 2000년 ~ 현재 서강대학교 컴퓨터학과 박사 과정. 관심 분야는 자연어 처리, 정보 검색, 지능형 에이전트

강 기 선
1998년 서강대학교 컴퓨터학과 학사. 현재 서강대학교 컴퓨터학과 석사과정. 관심 분야는 시스템 요구공학, 변경관리, 시스템 모델링.

김 재 선
1998년 서강대학교 컴퓨터학과 학사. 현재 서강대학교 컴퓨터학과 석사과정. 관심 분야는 시스템 요구공학, 에이전트, 분산시스템.

박 수 용
정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 1 호 참조

서 정 연
정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 1 호 참조