한국산업정보학회논문지 제13권 제2호(2008. 6)

구술문서 자료분석을 위한 정보검색기술의 응용†

(Information Technology Application for Oral Document Analysis)

박순철*, 함한희**
(Soon-Cheol Park, Han-Hee Hahm)

요 약 본 연구는 정보검색기술을 응용해서 구술문서 자료를 효율적으로 분석하는 시스템 개발을 목적으로 한다. 여기서 사용된 기술은 용어검색, 문서요약기술, 클러스터링기술, 문서분류기술, 주제추적기술 등이 있다. 본 연구를 위해서 전북지역에서 채록한 구술자료를 이용하였다. 구술문서 구조의 특성을 반영하면서 분석의 단위를 정하고 내용의 자동분류 및 분류체계에 따른 분류도 시도하였다. 특히 주제를 추적하면서 순서에 따라서 검색해가는 기술은 세계적으로도 아직 연구단계에 있던 것을 실제로 구현하였다. 이러한 5가지의 검색기술이 한 시스템에서 통합적으로 처리될 수 있다는 것도 이 연구가 이룬 성과이다. 이 연구의 기대효과는 구술문서 분석의 신뢰성·타당성·효용성을 높여서 구술문화연구에도 큰 기여를 할 것으로 기대된다.

핵심주제어 : 구술, 정보검색, 요약, 클러스터링, 문서분류, 주제추적

Abstract The purpose of this paper is to develop an analytical methodology of oral documents by the application of. Information Technologies. This system consists of the key word search, contents summary, clustering, classification & topic tracing of the contents. The integrated model of the five levels of retrieval technologies can be exhaustively used in the analysis of oral documents, which were collected as oral history of five men and women in the area of North Jeolla. Of the five methods topic tracing is the most pioneering accomplishment both home and abroad .In final this research will shed light on the methodological and theoretical studies of oral history and culture.

Key Words : Oral, Information Retrieval, Summarization, Clustering, Topic Tracking

1. 서 론

이 연구는 정보검색기술을 응용해서 구술문서 자료를 분석하는데 초점을 두고 있다.

최근 들어서 세계 각국은 과거의 문화를 보존하기 위하여 여러 방면(9 · 11사태 증언, 베트남전쟁 증언, 등)으로 노력을 기울이고 있다[1,2]. 특히 역사를 증언할 수 있는 사람들의 생생한 언어를 수집하여 이용·분석하고 있다. 국내에서도 이러한 경향에 동참해서 중요한 사건(4 · 3사태 증언, 종군위안부, 20세기민중생활사연구, 등)과 사람들을 대상으로 구술자료를 수집하고 있다[3-6]. 이처럼 구술자료 수집은 국내외에서 새로운 역사와 문화연구의 한 방법으로 자리를 잡아가고 있다[7,8]. 그럼에도 불구하고 국내에서는 구술자료를 본격적으로 이용·분석하는 방법에 대해서는 이렇다 할 논의가 진전되지 못하고 있다. 이러한 시점에서 본 연구는 정보검색기술을 구술문서에 응용해 봄으로 해서

† 본 연구는 2006년 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-321-A00012).
* 전북대학교 전자정보공학부 교수
** 전북대학교 고고문화인류학과 교수

이 분야에 새로운 방법을 제시해 보려고 한다.

정보검색기술이 구술문서 분석에 이용될 수 있는 이유는 구술이 갖는 특수성 때문이다. 구술은 문자와 다르게 장황하고, 체계적이지 못하며 동어반복적이고, 첨가적이어서 문자를 대할 때와는 다른 분석틀이 요구된다. 또 구술은 한 문서 안에서 앞뒤의 내용이 논리적이고 종속적인 관계를 유지하는 문자와 달리 언제라도 새로운 사실이 첨가되고 내용이 집합적이라는 특징이 있다[9]. 그러므로 연구자가 하나의 구술 텍스트에 들어있는 내용을 빠짐없이 그리고 등가적으로 처리하는 일은 쉽지 않다. 오히려 선입견을 가지고 임의적인 선택에 의하여 구술텍스트를 분석하는 경우가 많다. 그러나 컴퓨터는 입력된 모든 정보를 등가적으로 처리하기 때문에 자료처리에 있어서 신뢰성을 높일 수 있다. 또한, 사람의 인지능력은 제한적이어서 때로는 미세한 부분을 놓치는 경우도 적지 않다. 컴퓨터는 방대한 양을 동시에 처리하면서도 인간의 능력이 미치지 못하는 구석구석까지 숨어있는 내용을 찾아준다. 마침내는 구술내용을 중심으로 해서 문화의 새로운 패턴을 찾아낼 수 있는데 까지 발전하게 된다. 그러므로 컴퓨터에 의한 정보검색기술은 구술문서 분석의 유용성·신뢰성·타당성을 높여줄 수 있다.

본 연구에서는 5가지의 통합적인 정보검색기술을 구술문서 분석에 적용한다. (1) 용어검색[10] (2) 내용 요약[11-14] (3) 문서 클러스터링[15,16] (4) 분류[17] (5) 주제추적[18] 등이다. 이러한 통합적인 분석시스템을 통해서 구술언어가 문자언어보다 인간생활세계에 밀착되어서 풍부하고 중요한 정보를 전달하고 있다는 사실을 찾을 수 있다.

본 논문은 서론에 이어서 2장 구술문서의 구조와 특징. 3장 구술문서 분석시스템, 4장 시스템구현, 5장 결론과 향후과제로 구성된다.

2. 구술문서 구조 및 특징

본 연구를 위해서 사용된 구술자료는 2006-2007년 동안 전라북도 지역에서 수집한 구술자료이다(구술자료 수집에 참가한 전북대학교 연구원- 함한희, 강경표, 최우람, 주용기 등). 구체적인 설명을 위해서 이 자료 가운데 '여자선장 김순자(가명)의 일생'을 선택하여 예로써 설명하고자 한다. 이 구술문서는 하나의 일반적인 형태를 보여주고 있어서 구술문서의 구조와 특징을 설명하는데 충분하다.

연구자들이 수집한 구술문서는 면담자와 구술자가 일대일로 대화하는 장면을 그대로 보여주고 있는 것이 특징이다. 또 다른 특징은 구술자가 자유롭게 그리고 자연스럽게 이야기를 전개해 나가고 있다는 점에서 생활언어가 생생하게 살아있다[9]. 생활언어가 그렇듯이 논리적인 내용을 갖추거나 일정한 형식이 있는 것은 아니다. 그러므로 본 연구의 구술문서는 구술언어의 특성이 비교적 잘 드러나고 있다.

2.1 내용의 다양성

구술문서는 대체로 내용이 다양하다. 또 다양한 내용이 체계적으로 정리되지 않은 상태의 문서이다. 구술자는 일정한 주제를 가지고 이야기를 하지만, 구술의 과정에서 잡다한 이야기 꾸러미들이 삽입된다. 내용상 가지치기가 많다고 하더라도 그것들이 중요한 정보가 된다. 따라서 주제에서 벗어난 내용이라고 해서 소홀히 다룰 수는 없다. 구술문서에서는 어떤 내용이라도 빠뜨리지 않아야 하기에 면밀한 분석이 필요하다[6]. 하나의 구술문서 안에 들어있는 내용이 얼마나 다양한가를 보여주는 전형적인 예를 김순자(가명)의 이야기에서 찾으면 (예 1)과 같다.

인사 - 가족이야기1 - 생선장사 - 배구입 - 자녀교육 -가족이야기2 - 자수성가 - 결혼이야기 - IMF시기 - 장사이야기

(예 1) 김순자(가명)의 이야기

2.2 문단의 비연속성

구술문서는 면담자와 구술자 사이의 대화체 형식으로 구성된다. 대화체 문서에서는 면담자의 질문이나 코멘트 그리고 이에 대한 구술자의 답이나 자발적인 이야기가 하나의 문단으로 구성된다. 그러므로 분석을 위해서는 하나의 단위를 면담자와 구술자 사이에서 오고 간 한 뭉치의 이야기로 규

정한다. 아래의 (예 2)에서처럼 문단 1은 면담자의 질문과 구술자의 답이 하나의 단위가 되도록 묶는다. 문단 2 역시 마찬가지이다. 그런데 예문에서와 같이 문단 1과 문단 2는 연속해서 일어나는 이야기임에도 불구하고 실제의 내용은 다르다. 문단 1은 배를 구입한 이야기이고, 문단 2는 동생이야기이다. 연속해서 있는 문단의 내용이 비연속적이기 때문에 면담자와 구술자 사이에 오간 한 번의 이야기를 한 문단으로 취급할 수밖에 없다. 여기에서는 문단 단위의 분석이 필수적이다.

문단 1 면담자: 섬에 다니실 때는 배로 다니셨어요?
제보자: 여객선으로 다닐 때도 있고,, 인자 알으니까 주민들 배를 얻어 타고 갈 때도 있고. 그렇게 하다보니까는 배가 급선무더라구요. 그 뒤로부터 빚 얻어갔고 장만했죠.

문단 2 면담자: 네. 그게 언제쯤? 몇 살 때?

제보자: 배를 장만한 거는 한. 스물 한 여서 일곱 살. 인제 (동생이) 나보다 두 살 덜 먹었응게. 아버지 살아계실 때에는 아무것도 않고. 학교도. 공부도 안하고 그랬었는데. 철이 좀 빨리 들더라구요. 등치가 더 크니까.

(예 2) 면담자와 구술자의 말뭉치 예

2.3 응대, 접속어, 허사

구술문서는 정확한 내용을 파악하기 위해서는 정제의 과정이 필요하다. 대화체이기 때문에 구술자는 응수하는 말, 이어주는 말, 의미없는 발화 등이 자주 등장한다. 내용을 이해하는데 있어서 부차적인 언어습관을 제거하는 과정이 필요하고 이를 정보검색기술이 해결할 수 있다. 예문에서는 '그렇고말고'라는 응대어가 반복되고, '암', '근디' 등은 화자가 말을 이어가거나 머릿속의 생각을 정리할 때 사용되는 말들이 많이 나온다. 이러한 말들은 내용과는 직접 관련은 없다. 대화를 자연스럽게 이어가고자 할 때, 상대방의 의견을 확인할 때, 자신의 말을 가다듬고자 할 때 이러한 말들이 사용되고, 이것이 구술언어의 특징인 것이다. 정보검색기술을 이용해서 내용과 직접 관련이 적은 구어체적 습관에 의해 나타나는 응대어 및 접속어적인 성격을 띤 언어나 단어를 추출해 내는 기술이 적용되었다. (아래 예문 참조)

예문 1) 그렇고말고요. 아무렴, 그렇고말고. (여기에서는 그렇고 말고가 두 번 반복되었음)
예문 2) 암, 근디, 그건 그렇고, 허사

2.4 유사한 내용

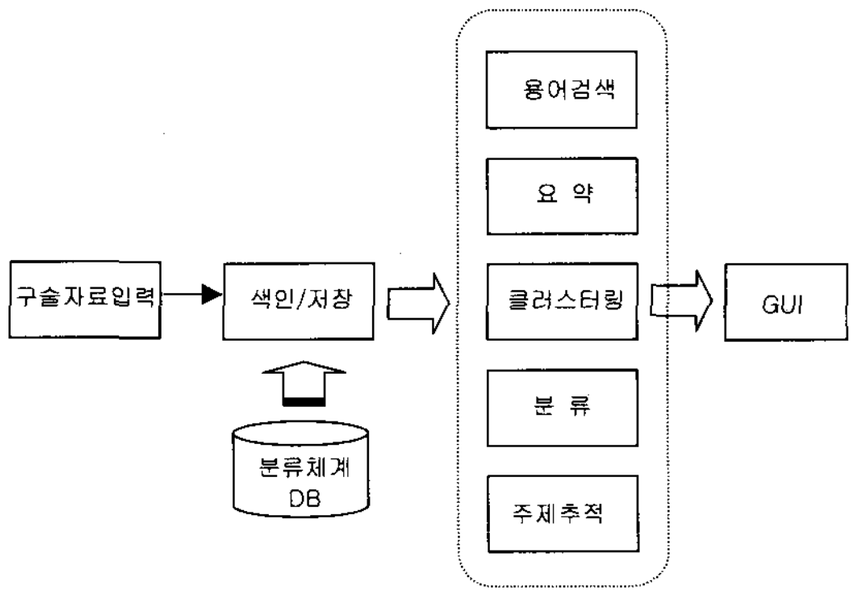
구술언어는 즉흥적이기 때문에 내용이 반복적이다. 앞에서 말한 내용이 다시 뒤에서 거듭 나오게 된다. 구술한 내용이 비슷할 경우에는 묶어주는 처리방법도 필요하다. 비체계적인 담화내용을 유사한 내용끼리 묶어서 분석을 하면 구술문서의 내용을 파악하는데 효율적이다. 이것은 문서의 자동분류 즉 클러스터링 기법과 분류에서 다시 설명하고자 한다. 2-1에서 지적한 바와 같이 구술의 내용이 다양하며 때로는 내용이 되풀이되는 경우가 많다. 가족이야기, 자녀교육 등이 중간 중간에 반복해서 나오므로 이러한 가족관련 주제를 자동으로 묶는 클러스터링기법을 적용시킬 수 있다. 이 기법을 이용하면 구술문서가 자동 분류되어서 분석이 용이해지는 이점이 있다.

2.5 주제추적

구술자는 자신이 하고 싶은 이야기를 의도한대로 전개해 나가기도 하고, 면담자의 질문에 따라서 말을 하기도 한다. 전자와 후자는 이야기 구성방법에서 차이가 나지만, 전체적인 내용에서는 큰 차이는 없다. 예를 들어서 구술자들이 사건이 일어난 시간의 순서나 논리의 순서에 따라서 체계적으로 이야기하는 것이 아니라는 뜻이다. 또한 면담자도 그러한 방식을 그대로 지키면서 말을 이끌어내기 힘들다. 따라서 연구자들이나 일반인들이 하나의 주제를 선정해서 구술자의 이야기를 순서대로 찾아가는 방법이 필요하며 이러한 방법이 구술문서의 내용을 분석하는데도 매우 유용하다. 위에서 인용한 구술문서의 예를 다시 들면, 여선장이 되기까지를 일관된 주제로 보고, 그 이야기를 순서대로 찾아갈 경우, 면담자가 현장에서 놓치거나 구술자의 의도가 잘 드러나지 않던 내용들도 주제추적과정에서 모두 드러나게 된다.

3. 구술문서 분석시스템

구술문서 분석시스템의 개략 구조는 (그림 1)과 같다. 그림에서 보이는 것처럼 시스템은 구술자료를 입력받아 자료에 포함되어있는 색인어를 구별하여 관련자료와 함께 저장한다. 이때 문서분류를 위한 분류체계의 용어들도 함께 입력된다. 이 시스템의 주요 기능은 용어검색, 요약, 클러스터링, 분류, 주제추적 등이 있다.



(그림 1) 시스템의 구조

주요 기능의 상세설명은 다음과 같다.

3.1 용어검색

용어검색은 구술자료 중 명사를 추출하여 색인된 색인어를 중심으로 이루어진다[10]. 검색에 대한 입력은 자연어처리가 가능하여 사용자가 원하는 문장을 임의로 입력하더라도 원하는 검색을 할 수 있도록 했다. 이러한 처리과정은 자연어로 입력받은 검색어를 파싱하여 원하는 색인어를 추출한 후, 추출된 색인어에 따라 저장시스템에서 관련자료를 찾아 보여주는 것이다. 기본적으로 검색어에 포함되어있는 색인어가 하나 이상일 경우에 'AND' 오퍼레이션을 하여 검색어에 포함되어있는 색인어를 모두 포함하는 경우의 자료만 찾게 된다. 본 연구에서 사용하는 검색모델은 통계학적 모델로 용어의 가중치를 통계값으로 사용한다. 식(1)은 용어의 가중치를 구하는 계산식이다. 이 가중치는 검색 결과에 문서 순위를 정하는 것, 문서 요약, 문서 클러스터링 등에 사용이 된다. 가중치를 계산하는 방법은 일반적으로 Tf(term frequency)와 idf(inverse document frequency)를 이용한다 [19,20]. 그러나 본 연구에서는 가중치의 계산에 용어의 특성을 포함시켜 용어검색의 정확도를 높였다. 식 (1)에서 용어의 특성을 나타내는 P값은 문자 폰트, 문자의 크기, 문자의 타입에 따라 임의로 정하여졌다. 식 (1)에서 가중치 W_{ij}는 j문서에 나타나는 i번째 용어의 가중치이다.

$$W_{ij} = Tf_{ij} \cdot idf(w_{ij}) \cdot P(w_{ij}) \qquad (1)$$

여기서,

$$Tf_{ij} = \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 \times \frac{doclen_j}{avgdoclen}},$$

$$idf(w_{ij}) = \log\left(\frac{N - df_{ij} + 0.5}{df_{ij} + 0.5}\right),$$

$$P(w_{ij}) = \begin{cases} 2.0: & high \\ 1.5: & important \\ 1: & others \end{cases}.$$

3.2 문서요약

문서요약은 문서의 내용을 중복없이 간략하게 요약하는 기술로서 전체의 내용을 읽기 전에 그 내용을 파악할 수 있도록 한다[11-14]. 이 기술은 통계학적 접근과 의미적 접근 방법에 의해서 연구되고 있다. 본 논문에서는 상대적으로 구현이 쉽고 적용이 가능한 통계학적인 방법을 사용하였다.

문서 요약은 색인어의 가중치를 중심으로 각 문장의 중요도를 결정한다. 즉 각 문장에 포함되어 있는 색인어들의 가중치를 합한 값에 의해서 문장의 중요도를 결정한다. 문서 요약 알고리즘은 식 (2)와 같다.

$$\arg\max^{k} \frac{\sum_{i=1}^{|passage|} Tf_{ij} \cdot idf(w_{ij}) \cdot P(w_{ij})}{|passage|} \qquad (2)$$

여기서,

$$Tf_{ij} = \frac{tf_{ij}}{tf_{ij} + 2},$$

$$idf(w_{ij}) = \max\left(M, \log\frac{N}{df_{ij}}\right),$$

$$P(w_{ij}) = \begin{cases} 2.0: & high \\ 1.5: & important \\ 1: & others \end{cases}.$$

식 (2)에서 사용된 Tf, idf, P값의 정의는 식(1)에서의 정의와 유사하다.

또한 본 시스템의 문서 요약의 특징은 Maximal Marginal Relevance (MMR)를 추가한 것이다 [12,13]. MMR의 특징은 문서 요약에서 적절한 문장을 선택함으로써 중복성을 감소시키는 것이다. 식 (3)은 본 시스템에서 사용된 MMR 알고리즘이다. 식 (3)에서 W값은 문장단위의 가중치를 의미한다.

$$\arg\max^{k}\left[\mathbf{W}_{passage_{new}} - \lambda \cdot \max sim(passage_{new}, passage_{olds})\right] \quad (3)$$

3.3 문서 클러스터링

본 시스템은 문서 검색 결과에 대하여 유사한 것끼리 모으는 문서 클러스터링 기능을 포함한다 [15,16]. 문서클러스터링기술은 문서 내의 유사한 내용을 자동 분류하여 사용자들에게 전체 자료의 구조를 쉽게 파악할 수 있도록 하며 분류된 각 내용을 집중 탐색할 수 있도록 도와준다. 이 기술은 K-means, MST, 등이 있으며 본 논문에서는 수정된 K-means 알고리즘을 이용하여 분류 성능을 향상시켰다[21,22].

문서 클러스터링 방법은 수정 K-Means 알고리즘이다. 일반 K-Means 알고리즘은 클러스터의 수가 정적인데 반하여, 본 시스템의 알고리즘은 그 수가 가변적이다. 즉 클러스터링하려고 하는 문서의 집단의 밀집된 정도에 따라 그 수가 동적으로 변하도록 했다.

K-Means 알고리즘은 이해하기 쉽고 구현이 간단하다. (그림 2)는 시스템에서 사용된 K-Means 알고리즘이다.

1. k 값을 선택한다.
2. 문서집합 d 에서 k 개의 proto-centroids 를 선택한다.
3. 거리 $dist(d_i, c_j)$를 계산한다.
4. 문서 d_i 를 다음 기준에 따라 클러스터 G_{c_j} 에 할당한다.
   a. $\arg\min_{\substack{i=1,n \\ j=1,k}} dist(d_i, c_j)$
   b. $d_i \underset{i=1,n}{\in} G_{c_j}$ $if\ dist(d_i, c_j) < dist(d_i, c_l)\ for\ all\ l = 1, 2, \cdots, k \quad l \neq j$
5. centroids $\vec{c}_j = \frac{1}{|c_j|}\sum_{l=1}^{|c_j|}\vec{d}_l$ 를 재계산 한다.
6. **if** $\max \delta(c_j, c_j^{new}) < \theta$ **then** 계산을 종료한다.
   else centroids $c_j = c_j^{new}$ 셋팅한 후, 3단계로 이동한다.

(그림 2) K-Means 알고리즘

3.4 문서분류

문서분류기술은 사용자가 정의한 분류에 따라 문서 또는 문서의 내용을 분류하는 기술이다. 본 연구에서는 종교를 중심으로 구술내용을 분류하여 서로 관계된 자료의 내용을 추출하였다. 기존 기법은 KNN, 코사인기법, 등이 있으나 본 연구에서는 구현이 간단한 코사인기법을 이용했다[17].

기존의 데이터 분류시스템들은 복잡하고 느리다. 이러한 단점을 해결하기 위하여 데이터 벡터가 유사도 측정을 위하여 Cosine Similarity를 이용한다. 자동 요약을 통하여 만들어진 요약내용을 이용하여, 속도가 빠르고 구현이 용이하다.

문서분류의 구체적인 알고리즘은 (그림 3) 과 같다.

1. 데이터와 분류체계 사이의 유사도를 계산한다.
2. 각각의 유사도를 계산하여 계산된 값이 가장 큰 카테고리로 문서를 분류한다.
3. 정확성 향상을 위하여 분류체계 내의 용어를 확장한다.

(그림 3) 문서분류 알고리즘

3.5 주제추적

문서내용추적기술은 한 가지 주제를 따라 원 자료의 구성순서와 동일하게 내용을 추적해 나가는 기술이다[18]. 이 기술을 적용함으로 다양한 내용의 구술 중 주제에 따라 전개되는 구술내용을 추출해 낸다. 이 기술에 대한 연구는 활발하게 진행되고 있으나 아직은 실제로 사용된 예는 없다.

주제추적은 전문가가 선택한 주제 혹은 사용자가 정한 임의의 주제에 따라 문서자료로부터 주제에 맞는 내용을 탐색하여 내용의 순서대로 추출하는 알고리즘이다. 주제추적의 구체적인 알고리즘은 (그림 4)와 같다.

1. 주제 T를 정한다.
2. 주제 T와 새로운 문단간의 유사도를 구한다.
3. **if** 유사도의 값이 일정값 이상이면, 그 문단을 주제 리스트에 포함시킨다.
   else 다음 문단으로 이동한다.
   문단이 선택될 때까지 step 3을 반복한다.
4. 주제 리스트에 있는 문단과 새로운 문단간의 유사도를 구한다.

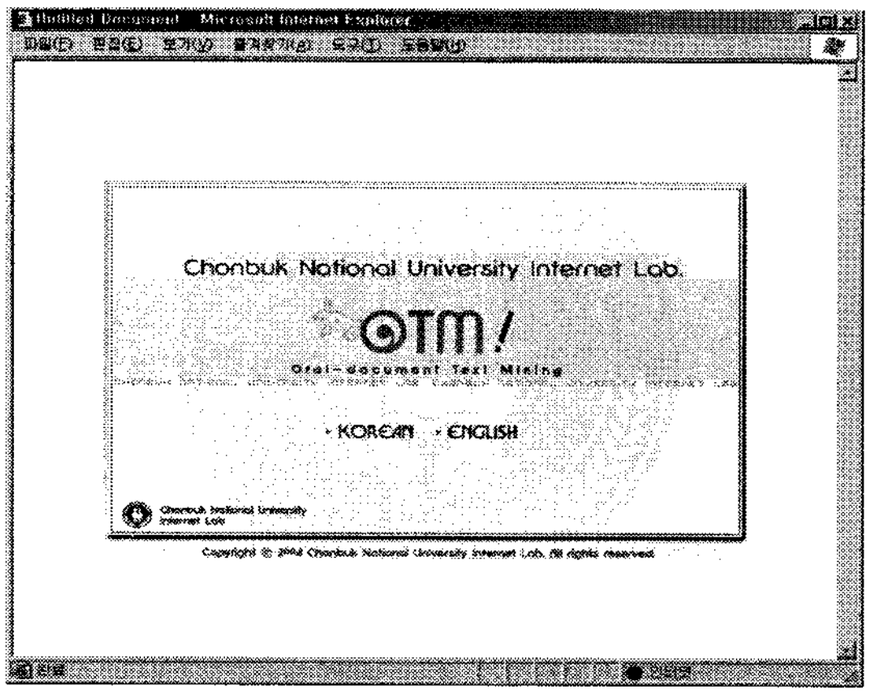
5. **if** 유사도의 값이 일정값 이상이면, 그 문단을 주제 리스트에 포함시킨다.
   else 다음 문단으로 이동한다.
6. 문서의 끝이 아니면 step 4를 반복한다.

(그림 4) 주제추적 알고리즘

4. 시스템 구현

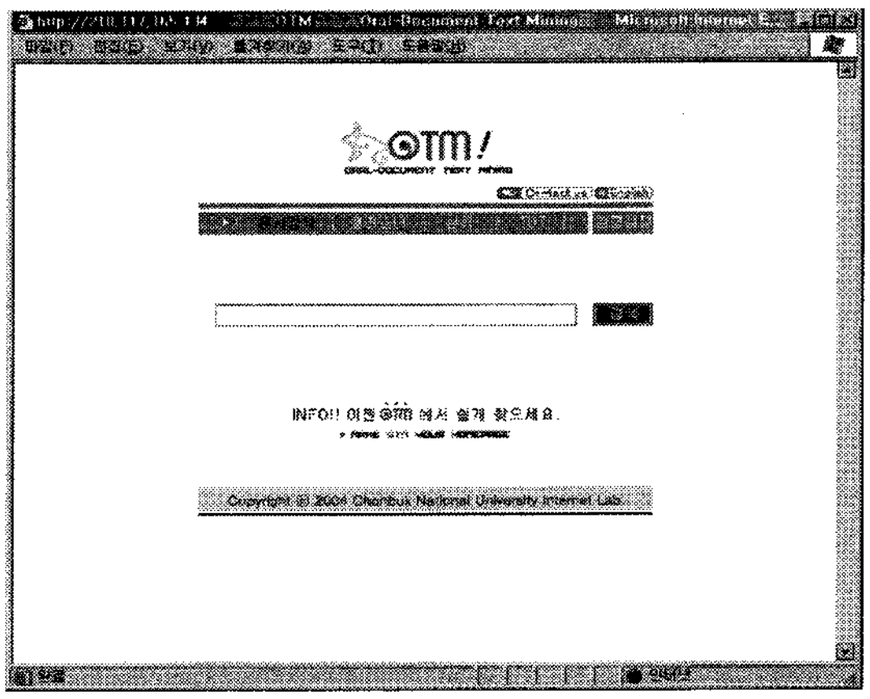
본 시스템은 웹상에서 작동할 수 있도록 구현하여 사용자가 언제 어디서나 쉽게 자료를 검색하고 분석할 수 있다. 구현된 시스템에 사용된 자료는 전라북도 지역에 거주하는 5인의 구술내용이다. 현재 시스템의 수행 속도는 실시간으로 동작된다.

본 장에서는 구현된 화면 결과를 중심으로 구현 시스템을 설명한다. 시스템의 첫 화면은 (그림 5)와 같다. 첫 화면에서 보이듯이 사용자는 입력된 자료에 따라 한국어 자료나 영문자료를 선택할 수 있다. 자료선택이 끝나면 (그림 6)의 화면으로 넘어간다.



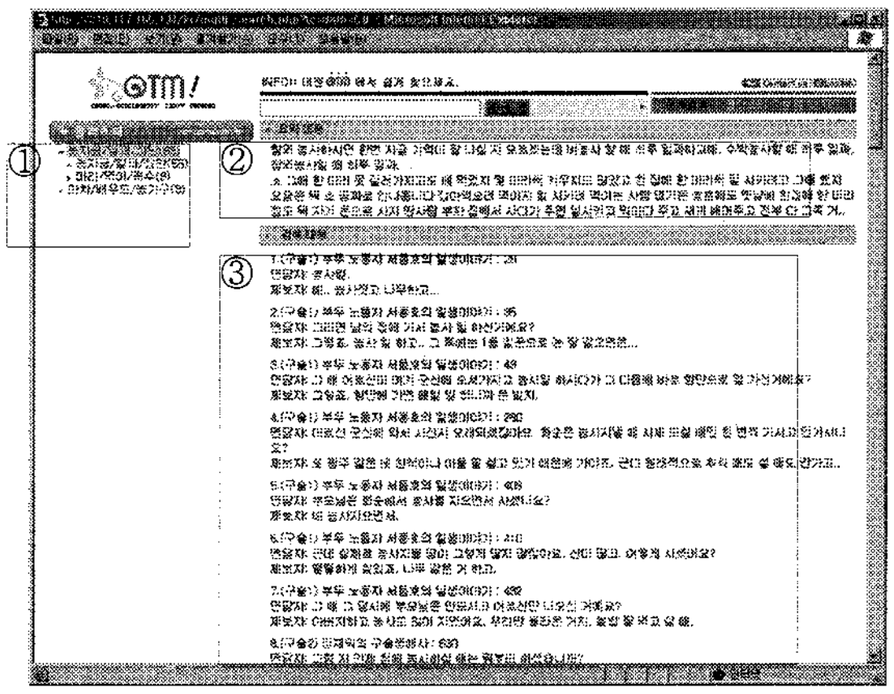
(그림 5) 초기 화면

(그림 6)는 시스템의 검색화면을 디폴트로 하여 원하는 기능화면을 선택할 수 있을 뿐만 아니라 환경설정 화면을 선택할 수 있다. 검색 창에서 질의어가 입력되면 바로 (그림 7)과 같은 검색 결과 화면이 나타난다.



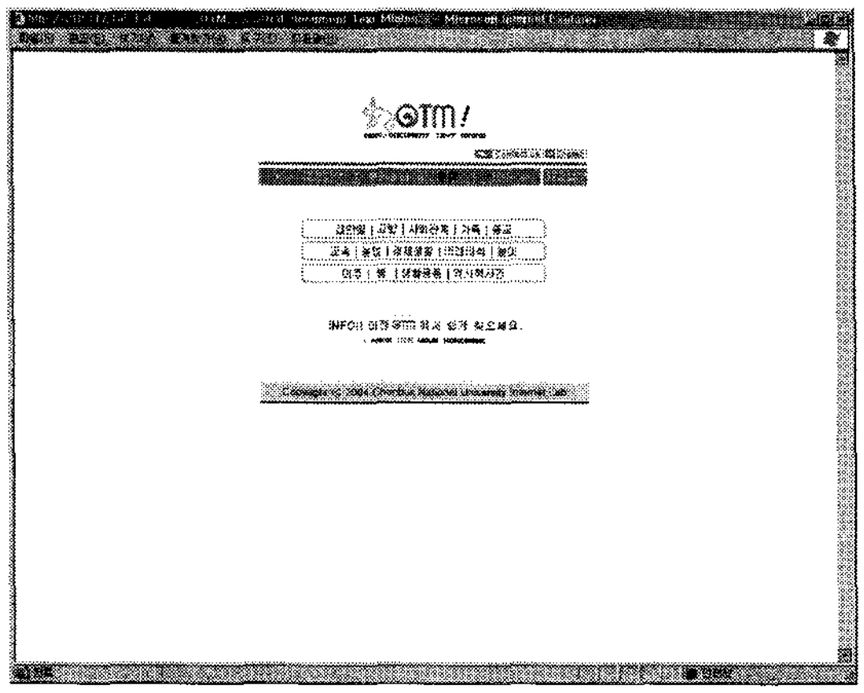
(그림 6) 실행 초기 화면

(그림 7) 검색결과 화면은 (그림 6)의 실행화면에서 검색용어를 '농사'로 했을 때 결과화면이다. 이 화면에서 ①번 창은 찾아진 결과에 대한 내용을 요약한 것이다. ②번 창은 검색된 자료를 순위별로 정렬하여 자료의 제목, 위치, 요약, 그리고 미리보기 기능 등을 포함한다. ③번 창은 검색된 자료들을 유사한 것끼리 클러스터링 하여 그 결과를 계층적 트리 구조로 가시화한 것이다.



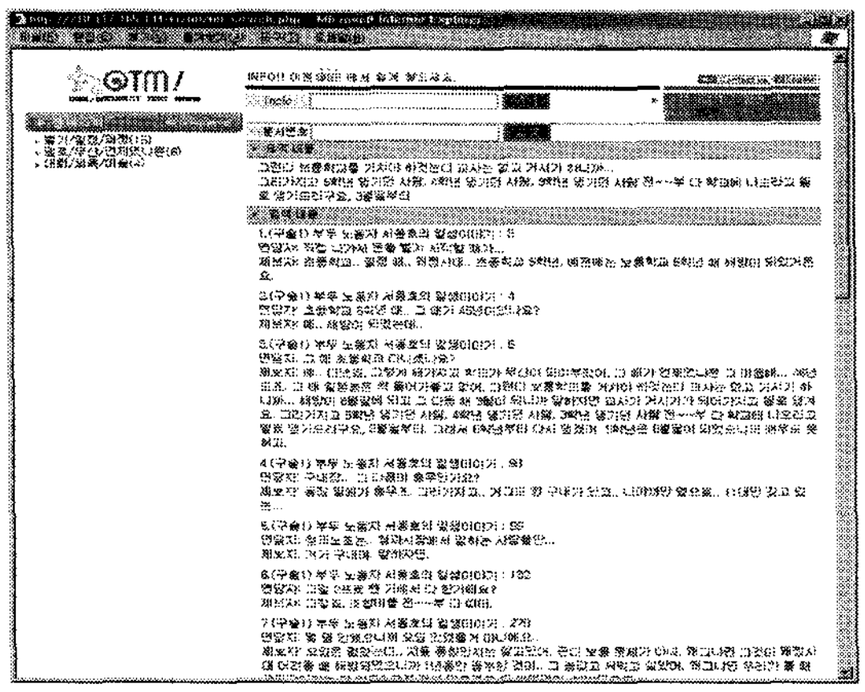
(그림 7) 검색결과화면

(그림 8) 자료분류화면으로 전문가에 의해서 분류된 항목에 따라 자료를 분류할 수 있게 한 화면이다. 입력된 데이터가 민중들의 일상생활에 관련된 것으로 여기서는 분류항목을 집안일, 고향, 사회관계, 등으로 했다.



(그림 8) 자료분류화면

(그림 9) 는 주제추적화면이다. '학교'라는 주제를 입력한 결과를 보인다. 결과내용을 자료에서 나타나는 순서대로 출력함으로써 주제의 내용이 어떻게 전개되는 지를 파악할 수 있다.



(그림 9) 주제추적화면

어떤 기능을 선택하든지 그 결과는 유사한 형태의 화면을 출력한다. 즉, 모든 화면의 왼쪽 창은 결과에 대한 클러스터정보, 오른쪽 상단은 결과에 대한 요약정보, 오른쪽 하단은 각 기능에 대한 결과자료를 출력한다.

5. 결론 및 향후 과제

정보와 통신의 혁명으로 인간의 말인 구술언어를 더욱 쉽게 보관하고 이용할 수 있게 되었다. 따라서 구술자료에 대한 중요성이 더욱 확산되고 있다. 그러나 아직까지도 이용과 분석차원의 연구가 답보상태에 머물고 있다. 이러한 점을 착안해서 본 연구를 진행하였다.

본 연구에서는 정보검색기술을 응용해서 구술문서 분석의 새로운 방법을 창출하였다. 그 결과 문서 속에 담긴 정보를 최대한 정확하고 쉽게 이용할 수 있게 되어서 분석의 신뢰도와 효용성을 높일 수 있다. 또한 정보검색기술이 구술자료 안에 들어있는 미세한 부분까지를 놓치지 않고 찾으며, 방대한 양을 한꺼번에 처리할 수 있기 때문에 지금까지 숨어있던 자료와 문서의 패턴을 보여주는 기능(문서요약, 문서클러스터링, 문서분류, 주제추적)을 제시하였다.

본 연구에서 제안하는 각 기능과 구술자료의 관계는 다음과 같다.

1. 용어검색은 사용자들에게 문서검색의 질과 효율을 높이도록 했다. 이 기능을 이용하여 사용자들은 문서의 전체 내용 속에 자신들이 원하는 주제내용을 쉽게 찾고 그 결과를 쉽게 분석할 수 있다. 그러나 앞으로 정확한 색인어 추출기와 효율적인 저장시스템의 연구가 필수적이다.
2. 문서요약은 통계적 요약방법에 MMR 기법을 적용하여 요약내용이 훨씬 더 함축적이고 분명하다. 이 방법은 사용자가 원하는 자료의 내용을 요약하여 제공한다. 흔히 구술내용이 방대하고 다양하기 때문에 요약방법은 많은 양의 자료를 신속하고도 용이하게 볼 수 있는 장점이 있다.
3. 문서 클러스터링은 검색결과를 포함하여 이 시스템에서 제공하는 모든 기능의 결과에 대한 클러스터링을 가시화하여 제공한다. 사용자는 클러스터링 구조를 통하여 구술자료가 가지고 있는 구조를 분명히 파악할 수 있다. 단순한 검색보다 훨씬 더 빠르고 정확하게 원하는 정보를 찾을 수 있을 뿐더러 내용이 포함하고 있는 숨어있는 패턴도 유추해낼 수 있는 도구로 사용될 수 있다.
4. 문서 분류는 구술자료의 분석에서 필수적인 분야이다. 현 시스템에서는 일반적인 수준의 분류체계를 이용하여 분류가 이루어진다. 향후 과제로 전문가에 의한 심도 있는 분류체계가 연구되면 본 시스템의 분류의 질도 향상될 것이다.
5. 본 논문에서 소개한 주제추적의 방법은 다양하

고 복잡한 내용과 구성을 가지고 있는 구술문서 자료에서 한 가지 주제를 가지고 내용을 순서대로 추출하도록 하는 것이다. 이러한 시도는 세계적으로 아직 연구단계에 있지만 본 연구에서 시험적으로 구술자료분석에 응용해 보았고 매우 유용할 것으로 생각된다.

이상의 특징을 가지는 구술문서 분석 시스템의 개발로 향후 구술언어 및 구술자료와 관련된 분야의 연구방법과 이론성립에 큰 기여를 할 수 있을 것으로 기대한다. 현재 개발이 미진한 몇 가지 분야가 향후 과제로 추가되면 본 시스템은 더욱 신뢰도·타당성· 효용성을 확장하게 될 것으로 보인다.

참 고 문 헌

[1] Columbia University Oral History Research Office, http:columbia.edu/cu/lweb/indiv/oral/

[2] The Vietnam Project, www.vietnma.ttu.edu

[3] 제주4·3연구소, 이제사 말햄수다, 한울, 1989.

[4] 제주4·3연구소, 무덤에서 살아나온 4·3 수형자들, 역사비평사, 2002.

[5] 한국정신대문제대책협의회, 강제로 끌려간 조선인 군위안부들 1-4, 한울, 1993-2001.

[6] 20세기민중생활사연구단, 한국민중구술열전 1-28권, 눈빛, 2005-2007.

[7] Dunaway, David K. and Willa K. Baum (eds.), *Oral History: An Interdisciplinary Anthology*, Nashville: American Association for State and Local History(AASLH), 1984.

[8] Mercier, L. & M. Buckendorf, *Using Oral History in Community History Projects*. The Oral History Association, 1992.

[9] 윤택림·함한희, 새로운 역사를 쓰기 위한 구술사 연구방법론, 아르케, 2006.

[10] Berry, M. W., and Murray Browne, *Understanding Search Engines*. University of Tennessee.

[11] Hand, Therese., "A Proposal for Task-Based Evaluation of Text Summarization Systems," In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 1997.

[12] Carbonell, Jaime. and Goldstein, Jade., "The use of MMR, diversity-based reranking for reordering documents and producing summaries," In Proceedings of ACM-SIGIR'98, Melbourne, Australia, August 1998, pp. 335-336.

[13] Mittal, Vibhu., Kantrowitz, Mark., Goldstein, Jade., and Carbonell, Jaime., "Selecting Text Spans for Document Summarizes: Heuristics and Matrics," In Proceedingsof the 16th National Conference on Artificial Intelligence, 1999, pp. 467-473.

[14] Goldstein, Jade., Kantrowitz, Mark., Mittal, Vibhu., and Carbonell, Jaime., "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," In Proceedings of ACM-SIGIR'99, Berkeley, CA, August 1999, pp. 121-128.

[15] Leuski and Allan, J., "Improving interactive retrieval by combining ranked lists and clustering," In Proceedings of RIAO'2000, April 2000, pp. 665-681.

[16] Torma, Markus., "Comparison Between Three Different Clustering Algorithms," Photogram metric Journal of Finland, Vol. 13, No. 2, Espoo 1993, pp. 85-95.

[17] Lam, W. and Ho, C.Y. "Using a Generalized Instance Set for Automatic Text Categorization," SIGIR'98, 1998, pp. 81-89.

[18] James Allen, *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers, Boston/Dordrecht/ London, 2002.

[19] Salton, G., and Buckley, C., "Term-weighting approaches in automatic text retrieval," Information Processing and Management, 24 (5), 1988, pp. 513-523.

[20] Jin, R., Faloutsos, C. and Hauptmann, A., "Meta-Scoring: Automatically Evaluating Term Weighting Schema in IR without Precision-Recall," The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, September 9-13, 2001.
[21] http://nlp.korea.ac.kr/~bewise/research/K-Means.pdf
[22] http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust5_bdy.html

박 순 철 (Soon-Cheol Park)

- 종신회원
- 1979년 2월 : 인하대학교 공과대학 (공학사)
- 1991년 12월 : (미국)루이지아나 주립대학 (전산학박사)
- 1991년-1993년 : 한국전자통신 연구원
- 1993년-현재 : 전북대학교 전자정보공학부 교수
- 관심분야 : 정보검색, 데이터마이닝, 디지털아카이브

함 한 희 (Han-Hee Hahm)

- 1975년 2월 : 서강대학교 사학과 (학사)
- 1990년 1월 : (미국)콜롬비아대학교 (인류학박사)
- 1990년-현재 : 전북대학교 고고문화인류학과 교수
- 관심분야 : 역사인류학, 문화아카이브