경기21서치 2.0 : 수치지도와 웹 공간을 융합한 지역지식 검색시스템†

Gyeonggi21Search 2.0 : Regional Knowledge Retrieval System using Numerical Map and the Web

윤성관*[1], 이용[2], 권용진[1]
Seong-kwan Yun*[1], Ryong Lee[2] and Yong-jin Kwon[1]
한국항공대학교 정보통신공학과[1], 삼성종합기술원[2]
{zikymi, yjkwon}@tikwon.kau.ac.kr[1], ryong.lee@samsung.com[2]

요 약

웹의 폭발적인 성장으로 다양한 형태의 지역관련 정보가 웹 공간에 포함되어 있으며, 기존의 지리정보시스템에서 제공하지 못한 실생활의 다양한 지역 정보를 얻을 수 있게 되었다. 하지만, 사용자가 지역 정보를 얻기 위해서는 현재의 키워드 기반의 웹 검색 엔진을 사용하여 얻은 다수의 검색 결과와 이를 맵과 관련시켜 정리해야 하는 불편함이 있다. 이러한 문제를 해결하기 위하여, 본 논문에서는 특정지역에 대한 정확한 지리정보를 갖고 있는 수치지도와 방대한 지역정보를 갖고 있는 웹 공간을 융합하여 특정지역과 관련된 지역정보를 효율적으로 제공할 수 있는 시스템인 『경기21서치 2.0』을 제안한다. 본 시스템에서는 웹을 통해 분석한 지역의 특징 및 지역 간의 의미적 관련성을 키워드를 통해 지역지식 네비게이터로 구성하고, 이를 수치지도에 기반한 맵 인터페이스와 연동하여 보다 효율적인 지역 웹 정보검색을 지원한다.

1. 서론

최근 인터넷 기술의 발전과 웹의 폭발적인 성장으로 웹을 통해 다양한 형태의 지역정보가 공유되고 있다. 이러한 지역정보에는 문화예술, 여행, 역사, 지역공동체 등의 다양한 범주의 정보를 포함하고 있다. 또한, 여행계획, 지역현황 분석 등의 현실공간에서의 사람들의 의사결정에 필요한 중요한 정보원으로써 사용되고 있다. 이는 기존의 지리정보시스템에서 이용되는 지도 및 지리정보를 통해 얻을 수 없는 광범위한 집단지성(Collective Intelligence)이다.

지금까지 지리정보시스템에서 사용되는 지리정보데이터베이스의 데이터는 지형공간에 관한 모든 정보, 즉 지리정보가 정형화된 데이터인데 반해, 웹에는 텍스트, 이미지, 동영상 등의 다양한 형태로 데이터가 존재하며, 지리정보뿐만 아니라 지역에 관한 문화·예술·경제·사회현상과 같은 지식정보도 가지고 있다. 또한, 기존의 정형화된 지리정보데이터베이스와 달리 간단한 링크구조로만 구성된 하이퍼텍스트 기반인 웹에서는 지리·지식정보의 발신, 생성, 공유가 자유롭지만 실공간과 관련시켜 정보를 검색하는 데는 어려움이 있다.

웹에서 지역정보를 검색하는 방법에는 카테고리 검색과 키워드 입력을 통한 검색이 있다. 카테고리 검색 방법은 사전에 전문 에디터에 의해 개념간의 의미관계를 고려하여 카테고리를 분류해 놓고, 사용자가 검색하고자 하는 영역의 카테고리를 선택함으로써 정보를 검색한다. 키워드 검색 방법은 사용자가 키워드, 즉 검색어를

† 본 연구는 경기도지역협력연구센터 (GRRC) 프로그램에 의해 한국항공대학교 차세대방송미디어기술 연구센터의 지원으로 수행되었음.

식섭 입력하는 방법이다. 전자의 경우는 사용자에게 원하는 정보를 제공하기 위해서 각각의 카테고리를 구성하는 과정에서 정보를 필요로 하는 사용자의 욕구를 고려한다는 측면이 있고, 후자의 경우는 광범위한 정보를 빠르게 검색해주는 장점이 있다.

한편, 지역정보에 포함되어 있는 지리적인 속성을 이용하여 지리정보시스템을 활용하는 검색 방법이 있다. 이러한 검색 방법은 웹 정보로부터 형태소분석을 통해 지리적인 속성을 가지는 단어들을 추출하여 해당 단어가 가지는 위치에 웹 정보를 매핑한다[1,2]. 사용자는 지리정보시스템을 통해 지역을 검색할 경우 지리정보와 함께 매핑된 웹 정보도 검색이 가능하다. 하지만, 대부분의 검색 결과는 주소, 전화번호 등의 간단한 정보만을 제공하고 있어 효용성이 낮다.[4]

본 논문에서는 특정지역에 대한 정확한 지리정보를 갖고 있는 수치지도와 방대한 지역정보를 갖고 있는 웹 공간을 융합하여 특정지역과 관련된 지역정보를 효율적으로 제공할 수 있는 시스템인 『경기21 서치 2.0』을 제안한다. 본 시스템은 웹을 통해 분석한 지역의 특징 및 지역 간의 의미적 관련성을 카테고리의 개념을 적용하여 키워드를 통해 지역지식 네비게이터로 구성했다. 또한, 이를 수치지도에 기반한 맵 인터페이스와 연동하여 보다 효율적인 지역 웹 정보검색을 지원한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 기술하고, 3장에서는 『경기21서치 2.0』 시스템 구현 내용을 작성하고, 4장에서는 본 시스템의 구성과 특징에 대해서 살펴본다. 5장에서는 결론 및 향후과제에 대해 알아본다.

2. 관련 연구

지리정보시스템은 디지털 형태로 변환된 지형 데이터를 이용하여 지도를 보여주는 것으로, 많은 종류의 데이터들이 중요한 지리적 형상을 가지고 있기 때문에 그 목적에 맞게 여러 가지 형태로 데이터를 가공하여 필요한 부분만 중점적으로

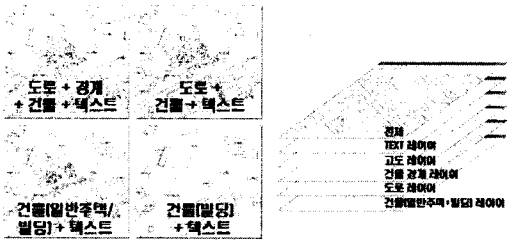

그림 2. 수치지도 계층 구조

보여줄 수 있다. 이러한 특성으로 교통, 일기예보, 인구예측, 문화재 관리, 도지계획 및 여행지의 위치정보 등 다양한 분야에 응용되고 있다[3].

수치지도는 현실공간의 객체를 수치로 표현하고 있는데 객체를 <그림2>처럼 레이어로 분류하고 있어 필요한 데이터만을 추출하여 사용할 수 있다. 특히 특정 지역에 관한 텍스트, 즉 지명, 유적지, 영소 등의 지역 명사를 자세히 포함하고 있어 특정 지역과 관련된 지리단어를 쉽게 구성할 수 있다.

웹 공간에서 수집한 지역정보로부터 새로운 지역정보·지식을 생성하기 위해서는 지역정보 사이의 상호관련성에 관한 분석이 필요하다. 주로 색인어를 이용한 방법을 통해 연관도를 계산하며, 데이터마이닝의 연관규칙(association rule)[5]과 행렬의 곱을 이용하여 유사도 행렬(similarity matrix)를 만드는 방법[6]이 있다.

연관규칙은 "트랜잭션(transaction)"이라는 "한 번에 함께"라는 정보를 이용하여 데이터를 분석하는 방법으로, 두 색인어 집합을 동시에 포함하는 문서의 수를 전체 문서의 수로 나누어준 지지도(support)와 두 색인어 집합을 동시에 포함하는 문서의 수를 한 색인어 집합을 포함하는 문서의 수로 나누어준 신뢰도(confidence)라는 두 척도를 이용하여 찾아낸다. 반면에 유사도 행렬은 "두 용어가 공기하는 문헌의 수"와 같은 문서에 색인어가 몇 번 출현하는지 빈도를 측정하여 행렬을 만들고, 생성된 행렬의 전치행렬을 구한 후, 두 행렬의 곱을 통해 두 색인어간의 연관성을 찾는 방법이다.

3. 웹 지역정보 추출

3.1 수치지도를 이용한 지리단어 추출

본 논문에서는 수치지도를 이용해 특정 지역을 대표할 수 있는 지리 단어 집합을 생성한다. 수치지도는 실제 지형을 수치로 표현한 지도로, 지형을 x, y좌표의 형태로 저장하고 있다. 도로나 강, 건물은 x, y 좌표 점들이 모여서 이루어진 선과 면으로 표현된다. 이러한 수치지도에는 현실 공간의 각 객체가 일괄적으로 정해진 표준코드로 분류되어 기록되어 있다. 각 코

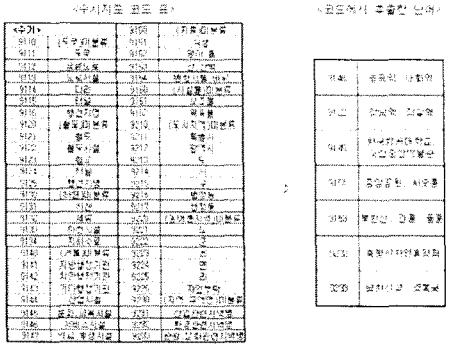

표 3.1 수치지도 코드 표와 추출 단어

드는 해당 객체가 강인지, 산인지, 도로인지 등을 상세히 나타내고 있다. 이러한 수치지도에는 지명이 포함되어 있고, 각 지명에 대한 코드도 따로 분류가 되어있어 코드를 기준으로 특정 지역과 관련된 단어를 추출할 수 있다. <표3.1>의 코드 분류 중에서 본 논문은 지역을 "경기도"로 한정하고 있으므로 이 지역과 관련된 지리 단어를 추출한다. 주로 관광·문화관련 지역명, 환경관련 지역명, 철도시설, 행정명 등과 같이 특정 지역을 대표할 수 있는 코드를 중심으로 단어를 이용한다.

3.2 특정 지역과 관련된 웹 문서 수집

일반적으로 주제기반 웹 수집은 전체 웹 정보에서 주어진 주제 또는 사용자의 관심과 관련한 일정한 세그먼트를 유지하고 관리하는데, 이러한 웹 수집 방법은 특정한 주제와 연관된 정도(degree of relevance)를 정의하여 이러한 모델을 바탕으로 대략적인 웹 페이지 집단(cluster)을 형성한다. 그리고 기계학습 알고리즘을 사용하여 자동적으로 생성한 웹 페이지 분류 규칙을 이용하여 웹 페이지 집단을 주제와의 관련성이 높도록 정제한다[7]. 이러한 주제기반 웹 수집 방법을 활용하면 특정 지역과 관련된 웹 문서를 수집할 수 있다. 하지만, 무작위로 웹 문서를 수집해서 해당 지명이 포함되어 있는지의 여부를 판단해야 하는데, 수집된 페이지가 특정 지역과 관련이 있는 페이지일 확률이 낮을 수밖에 없기 때문에 문서를 수집하는데 많은 시간이 소요된다.

본 논문에서는 "경기도"라는 특정 지역과 관련된 웹 문서를 수집하기 위해서 특정 지역과 관련된 문서의 자동 분류로써 웹 페이지를 수집해 오기 전에 해당 페이지가 "경기도"와 관련이 있다고 미리 판단하고 수집해 오는 방법을 사용한다. 수치지도로부터 추출한 지리단어를 이용하여 검색엔진에 질의로 보내, 결과 페이지로부터 URL을 추출한다. 검색엔진에서 분류해준 결과에 해당하는 문서들은 특정 지역과 관련 있는 문서라고 판단할 수 있기 때문이다. 그리고 이렇게 수집된 문서로부터 일정 범위 이내에 링크된 문서들은 대체로 초기 문서와 관련 있는 문서일 확률이 높다.[8] 그러므로 해당문서가 특정지역과 관련 여부를 문자열 비교를 할 필요가 없으므로 문서를 빠른 시간안에 수집할 수 있다.

3.3 공간정보원의 확대

『경기21서치 2.0』 시스템에서는 "지역관계"와 "지역 특징"을 지명단어와 비지명단어로 <그림3.1>과 같이 나타낸다. 이

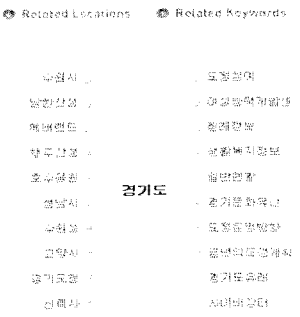

그림 3.1 키워드 인터페이스

러한 관계를 구성하기 위해서는 웹 문서로부터 형태소 분석기로 단어를 추출한 후, 특정지역과 관련된 단어를 찾아야 하는 문제가 있다. 이것도 앞의 특정지역과 관련된 웹 문서를 찾는 문제와 마찬가지로 기존의 자동으로 분류하는 방법들은 사전에 학습된 단어집합이나, 시소러스를 이용하는 등 한계가 있다.

하지만 본 논문에서는 "경기도"관련 단어 집합 생성에 검색 엔진을 이용하여 자동으로 분류하는 방법을 고려한다. 만약 해당 단어가 특정 지역과 관련이 있다면, 해당 단어를 이용하여 검색 할 경우, 결과 페이지에 특정 지역명이 포함되어 있을 확률이 높다. 실제로, "Google"에 "경기도"와 관련이 있는 "행주산성"이라는 쿼리를 주면 검색 결과 페이지 안에는 "경기도", "경기"와 같은 문자열이 포함되어 있다. 따라서 지역명이 포함되어 있는지 여부로 관련성을 판단할 수 있다.

이러한 방법으로 특정 지역과 관련된 단어 집합을 구성하고, 단어들을 수치지도에서 추출한 단어와 그 이외의 단어로 분류한다. 전자는 지리정보원으로써 지도위에 표현하여 지리정보를 제공하고, 후자는 해당 지역을 특징짓는 단어로써 비지리정보원으로 사용한다.

3.4 개념들 간의 관계 추출

『경기21서치 2.0』 시스템에서는 <그림3.1>과 같이 키워드 인터페이스를 구성하여 관련 있는 단어를 표시하게 된다. 이 키워

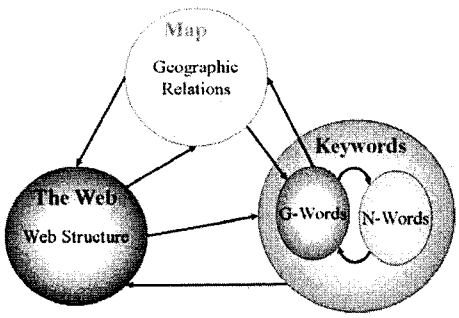

그림 3.2 공간사이의 관련성

드로 새로운 검색을 유도해 사용자가 원하는 정보를 효율적으로 제공하고, 예기치 못한 정보까지 새롭게 가공하여 보여준다.

단어들간의 연관성을 얻는 방법은, 웹 페이지의 분석으로부터 얻어지는 명사 집합을 지명(geoword, G)과 비지명(non-geoword, N)으로 나눈다. 그리고 수집된 웹 문서를 웹 정보공간(P-domain)으로 하여, 현실(지명)공간(G-domain), 비지명 공간(N-domain)을 구성하여, 이들 공간사이의 관련성을 추출한다. 이러한 관련성 추출은 새로운 지식을 생성한다. 이들 지식을 데이터 마이닝에서 사용하는 연관 규칙(Associative Rules)으로 표현하면 다음과 같이 표현할 수 있다.

*·지역의 특징: G->N**

*·지역간의 관련 : G->G**

*·지역 웹 페이지 : G->P**

이것들의 실제 계산에는 지역 관련 웹 페이지 G->P*으로부터 각각 특징과 지역 관련을 계산한다. <그림3.2>와 같이 공간

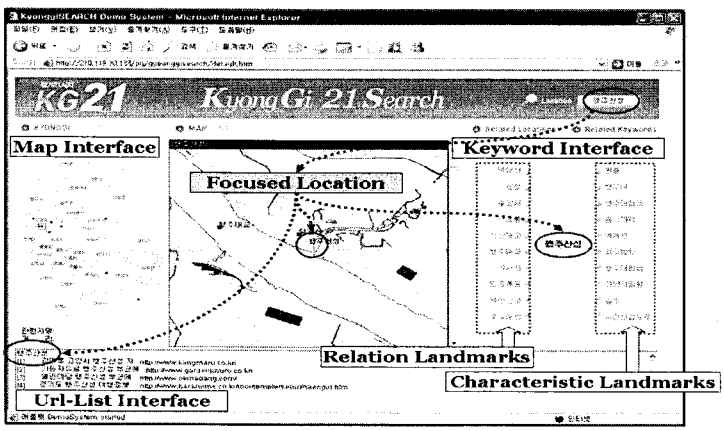

그림 4.1『경기21서치』 시스템의 구성

사이의 전체적인 관계를 생각함으로써 지역에 관한 복합적인 지식을 생성한다.

4. 『경기21서치 2.0』 시스템의 구성과 특징

『경기21서치 2.0』 시스템은 웹으로부터 추출된 지역적 개념간의 관계들을 이용해 새로운 공간정보원을 생성하고, 지도 표시를 통해 공간정보를 찾고, 관련 URL로부터 새로운 지역정보를 얻을 수 있는, 지리정보시스템과 웹 정보를 통합한 시스템이다.

4.1 시스템 구성

시스템은 <그림4.1>과 같이 맵 인터페이스, 키워드 인터페이스, URL 인터페이스로 구성되어 있다. 입력창에 키워드를 입력하면 "중심 키워드"가 되어 오른쪽의 키워드 인터페이스에 가장 연관성 높은 10개의 지명관련 키워드와 비지명(특성) 키워드가 표시된다. 중앙에는 맵 인터페이스가 있는데, 중심키워드가 "지명"이면 사용자에게 공간정보를 뿌려준다. URL 인터페이스는 하단에 있고 중심키워드와 연관성 있는 10개의 웹 URL이 나타난다.

4.2 시스템 특징

여행자가 기존의 지리정보시스템을 이용하여 여행지를 살펴보고 관련 정보를 얻기 위해 일반적으로 <그림4.2>와 같은 계획을 세울 것이다.

우선 여행자는 경기도의 문화재에는 어떤 것들이 있는지 찾기 위해 웹을 이용,

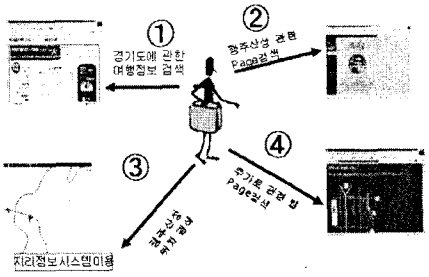

그림 4.2 기존 지리정보시스템과 웹을 이용한 여행정보 검색

경기와 관련된 여행정보를 검색한다. 그리고 행주산성이 있다는 것을 알게 되고 행주산성이 어떤 곳인지 알기 위해 웹을 이용해 행주산성과 관련된 페이지를 검색하고, 행주산성을 찾아가기 위해 지리정보시스템을 이용해서 위치를 검색한다. 그리고 행주산성에 대한 추가적인 정보를 웹에서 다시 검색한다.

본 시스템에선 이런 일련의 과정을 빠르고 쉽게 검색할 수 있다. 여행자가 "경기도"를 중심 키워드로 입력하면, 키워드 인터페이스에는 경기도와 가장 연관성이 높은 지역 명사 10개와 비지명(특성) 명사 10개가 표시된다. 여행자는 이들 키워드를 통해, 관심을 갖고 있는 "경기문화유산"이라는 단어를 선택하게 되고, 다시 "경기문화유산"이 키워드가 되어 가장 연관성이 높은 단어 20개가 나타난다. 여행자는 "경기도"-"경기문화유산"이라는 키워드를 통해 가장 관련 있는 지역 10개를 볼 수 있고, 그 중 "행주산성"을 선택 할

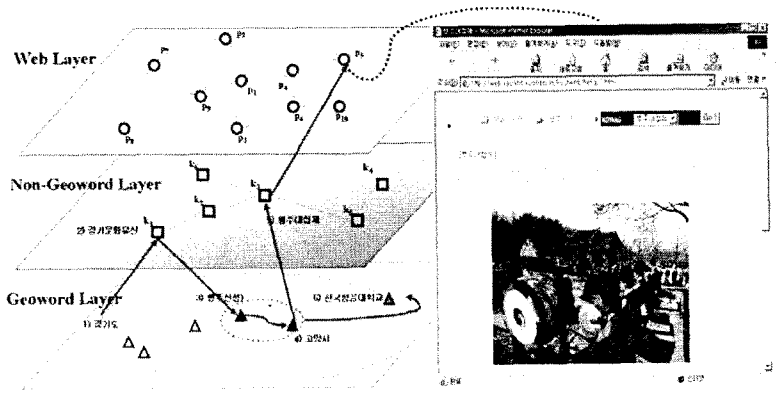

그림 4.3 『경기21서치2.0』의 지역정보 검색 흐름

수 있게 된다. "행주산성"이라는 지역 명사를 중심 키워드로 선택하면, 키워드 인터페이스에 20개의 단어가 다시 표시되고, 맵 인터페이스에서는 "행주산성"을 여행자에게 뿌려주며, 여행자는 행주산성의 공간정보를 알 수 있게 된다. 그리고 행주산성과 관련된 비지명 명사를 통해 "행주대첩제"라는 것이 있다는 것을 알게 되고, URL 인터페이스에서 "행주대첩제"와 가장 연관성이 높은 URL을 선택하여 정보를 얻을 수 있다. 이와 같이 『경기21서치2.0』은 맵과 다양한 웹 컨텐츠를 통합하여 여행자에게 효율적인 인터페이스를 제공한다. <그림4.3>은 이러한 흐름을 보여준다.

5. 결론 및 향후과제

본 논문에서 개발한 『경기21서치 2.0』 시스템은 수치지도의 지리정보와 웹 공간에 존재하는 정보를 융합하여, GUI를 통해 지역정보를 효율적으로 검색해주는 시스템이다. 이 시스템은 지도, 키워드, URL 리스트 등의 다양한 인터페이스가 상호 연동하면서 사용자에게 지역정보를 제공하고 있는 특징이 있다. 특히, 키워드 인터페이스의 경우 검색하고자 하는 지역을 중심으로 지명과 비지명(특성)으로 분류하여 제공함으로써, 사용자는 지리정보뿐만 아니라 지역과 관련된 문화, 예술, 역사에 관한 추가적인 정보도 함께 얻을 수 있게 된다. 또한, 검색하는 과정에서 의도하지 않은 지역정보도 획득이 가능한 시스템이다. 향후과제로는 위치를 가지는 지명단어 간의 거리를 새로운 인자로 고려하여 다양한 연관도 계산방법을 모색하고자 한다.

6. 참조

[1] K.A.V. Borges, A.H.F. Laender, C. B. Medeiros, A.S. Silva, and C.A. Davis Jr. "The web as a data source for spatial databases," in V Brazilian Symposium on GeoInformatics (GeoInfo 2003), Campos do Jordão (SP), 2003.

[3] Einat Amitay , Nadav Har'El , Ron Sivan , Aya Soffer, "Web-a-where: geotagging web content", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, July 25-29, 2004, Sheffield, United Kingdom

[2] Wenbo Zong , Dan Wu , Aixin Sun , Ee-Peng Lim , Dion Hoe-Lian Goh, "On assigning place names to geography related web pages", Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, June 07-11, 2005, Denver, CO, USA

[3] 장인성, 최혜옥, 이종훈, "Web-GIS를 위한 데이터 공유", 한국정보처리학회 추계 학술대회, 2001년

[4] Christopher B. Jones , R. Purves , A. Ruas , M. Sanderson , M. Sester , M. van Kreveld , R. Weibel, "Spatial information retrieval and geographical ontologies an overview of the SPIRIT project", ACM SIGIR conference, August 11-15, 2002, Tampere, Finland

[5] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Dataminig: A peformance perspective", IEEE Transactions on Knowledge and Data Engineering, 5 (6), 12, 1993

[6] Baeza-Yates, Ribeiro-Neto "Modern Information Retrieval", Addison-wesley, 1999

[7] 서혜성, 최영수, 노상욱, 최경희, 정기현, "연관도를 계산하는 자동화된 주제 기반 웹 수집기", 인터넷정보과학회논문지, 2006

[8] T. Tezuka, R. Lee, H.Takakura, and Y. Kambayashi, "Integrated Model and Implementation of a Region-Specic Search," 3rd IRC Int. Conf. on Internet Information Retrieval, pp. 243-248, 2003.