韓國 OA學會 論文誌
第7卷 第4號, 2002. 12.

2002-7-4-2-2

# 역사적 기록 문서에서 효율적인 유사도 및 클러스터링 측정에 관한 연구

한 광 덕*

# A Study on the efficiency of similarity and clustering measure in Historical Writing Document

Kwang-duk Han*

## 요 약

Web상에 있는 문서들이 다양하고, 복잡 그리고 대형화함에 따라 문서의 표현과 전달체계에서도 많은 변화가 요구되고 있다. 조선왕조실록문서(Annal of The chosun Dynasty)는 역사적 사실을 연구하는 데 중요한 문서이고, CD-ROM으로도 출판되었다. 그러나 문서의 접근 방법에 대해 검색의 단순성 그리고 내용 기반(content-based)으로 구성되었기 때문에 문서의 구성요소들 간의 사건연관(event-relationship)를 엮어주는 데는 어려운 점이 많다. 따라서 본 논문에서는 조선왕조실록 문서들 간의 효율적이고, 적절한 유사성 및 클러스터링 방법을 실험하여 문서들간의 사건연관을 찾아내도록 연구했다. 연구 방법으로는 조선왕조실록 문서들간의 유사도 방법들을 시뮬레이션하여 역사적 기록문서에 가장 적합한 유사도 방법을 찾아내고, 유사도 확률에 따라 그 문서들을 클러스터링 하였다. 평가결과 클러스터링을 한 문서들을 실제 확인해본 결과 사실과 거의 같다는 것이 증명되었다.

## Abstract

It expected a lot of changes in mass media and documentation expression as documents on web are getting diverse, complex and massive. An Annals of The Chosun Dynasty is a very important document used for researching historical facts and is published as CD-Rom. However, The CD-Rom was composed as content-based and using simple search method, therefore it's very difficult to make determine event-relationship between documents factors. Because of that, we studied to discover event-relationship between

* 상지 영서 대학교 컴퓨터 정보 기술과 조교수

논문접수 : 2002. 9. 17
심사완료 : 2002. 12. 1

documents through clustering and efficient similarity method among Annals of The Chosun Dynasty. For the research method, we discovered the best similarity method for historical written documents through simulation similarity measures of Annals of The Chosun Dynasty documents. Then we did simulation-clustering documents based on similarity probability.In evaluation of the clustered documents, the results were the same as when manually figured.

# Ⅰ. 서론

Web에 있는 다양한 문서들이 복잡하고, 대형화함에 따라 문서의 표현과 전달 체계에서도 많은 변화가 요구되고 있다. 문서 내에 기록되어 있는 정보 시스템의 자료는 두 가지 중요한 영역으로 구성된다. 첫째는 축적, 검색될 문서 또는 자료이고, 두 번째로 질의 혹은 정보 요구의 표현이다. 정보검색의 관점에서 볼 때 중요한 문제는 정보 요구를 어떻게 표현하고 , 그 요구에 부합하는 문서를 어떻게 찾아내느냐 하는 것이다. 이러한 과정은 자료가 어떠한 형태로 축적되느냐에 따라 영향을 받는다.[1] 이러한 조선왕조실록을 CD-ROM으로 개발되어 조선왕조실록을 연구하는데 능률과 효율을 크게 높이게 되었다. 따라서 정치사, 경제사, 사회사, 풍속사, 제도사, 문화사, 예술사 및 천문학, 지리학, 동식물학, 약학, 의학 등 각 분야별 분류사의 연구 및 거시적(巨視的)이며 통사적(通史的)인 역사 연구도 보다 능률적인 방법으로 접근할 수 있다. 현재 조선 왕조 실록 CD-ROM은 내용 기반(content-based)에 의한 full-text 단순 검색. 그리고 연대 목차, 분류 색인 검색의 단일성, 단순한 질의, 멀티미디어 문서로 표현할 수 있는 융통성 부족 등 아직 해결해야할 문제들이 있다. 클러스터링하는 방법중에는 유사도행렬에 의한 클러스터링 방법과 자기발견적 클러스터링으로 구분되어 클러스터링을 하였다. 그러나 역사적 기록문서의 특징을 살려서 유사도행렬에 의한 클러스터링에 자기 발견적 클러스터링을 혼합하는 방법을 제시한다.

따라서 본 논문에서는 역사적 기록 문서를 대상으로 단순한 질의뿐만이 아니라, 전문가가 제시한 정보의 요구를 충족시켜 줄 수 있는 event-relation 구축하기 위한 문서와 문서간의 유사성 측정 기법 및 클러스터링 시스템을 제안 및 역사적 기록 문서를 표현한다. 즉, 역사적 기록 문서에 나타난 어떤 사건은 그 사건 전에 발생했던 여러 사건의 영향을 받고, 그 후에 일어난 여러 사건의 영향을 끼침에 관련을 주어, 그 관련된 연관성을 묶음으로써 사건의 흐름을 관련 질 수 있는 클러스터링 구축에 관한 연구를 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련된 연구들을 설명한다. 3장에서는 시스템의 구조를 보이고 유사도 측정 및 클러스터링 기법에 대해 설명한다. 4장에서는 실험을 통해 유사도 측정의 정확도와 효율성, 그리고 클러스터링을 구현한다. 5장에서는 앞장에서 제시한 유사도 측정 기법, 클러스터링 기법과 실제 사항과의 비교 및 결과 분석을 한다. 6장에서는 결론을 내리고 향후 연구과제를 제시한다.

# Ⅱ. 관련 연구

## 1. 역사적 기록문서의 분류

### 1.1 시간 중심 기록문서

역사적 기록문서에서 가장 큰 특징을 가지고 있는 틀은 시간의 연속성이다. 신문, 방송 등을 예로 든다면 시간 중심적으로 발행이 현재까지 존속되어 있고, 시간 중심의 기본으로 그 내용은 사건이 기술되어 있어, 시간 중심 중심으로 기록한 문서종류는 편년체, 기전체, 기사본말체, 신문체(현재 일간지), 고려사등이 있다.

시간 중심 기록문서의 특징으로는 시간을 축으로 event들을 기술했고, 이는 시간대별, 목차별로 나열되어 있다. 대표적인 기록 문서로서는 조선왕조실록이 있다.

### 1.2 사건 중심 기록문서

외형상으로는 시간 중심 형식으로 구조화했지만 사건의 중대성을 감안하여 사건의 내용을 자세하게 기술한 것으로 역사 문서에서는 "비변사등록"이 그 대표적인 예이다. 사건 중심 기록문서의 특징으로는 시간 중심 기록문서사 흡사하나 사건의 내용을 상세히 기술하고, 서술 기법에는 육하원칙에 기본을 두었다. 대표적인 기록 문서로서는 비변사등록이 있다.

### 1.3 동일 event에 대한 시간 중심 기록문서 /사건 중심 기록문서의 통합

시간 중심 기록문서는 사건의 내용이 기본적인 것들만 기술되어 있지만, 사건 중심적 문서에서는 사건의 내용이 상세하게 기술되어있어서, 전문가의 특정 질의 요구에 적

절히 제공해 줄 수 있다. 따라서 검색에서는 사건 중심의 문서를 기본으로 하여 사건 중심의 문서와 유사도를 비교한다면 좀 더 확실한 정보를 적절한 응답을 할 수 있어, 동일 사건에 대해서 역사적 기록 문서들을 서로 연결시킨다면, 검색 효율 증가와 더 많은 정보를 얻을 수 있는 특징을 가질 수 있다.

### 2. 조선 왕조 실록

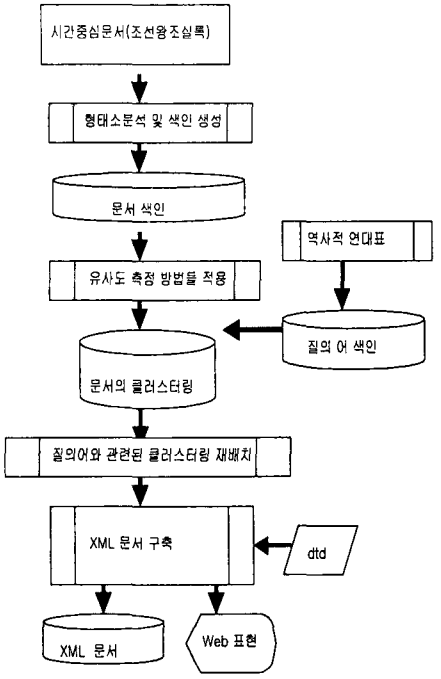
조선왕조실록은 조선 태조에서 철종까지 472년 간 역사적 사실을 각 왕별로 기록한 편년체 사서(編年體史書)로써 국보 제 151호 이고, 기록 시대는 1413년(태종 13)~1865년(고종 2)까지 각 왕별로 기록된 총 1,893권 888책로 구성되어 있다.

### 3. 조선 왕조 실록 CD-ROM[5]

조선왕조실록 국역작업이 1993년 말, 세종대왕기념사업회 및 민족 문화추진회에 의해 완료됨에 따라. 1994년 4월 문화체육부, 교육부 및 양 번역기관과 서울시스템이 합의하여, 1994년 5월 한국학 데이터베이스 연구소를 설립하여 개발을 시작, 학술용 확장 한자 부호계 및 확장한자 서체개발, 한글, 한자 full-text 검색엔진 개발 그리고 KS 표준한자인 4,888자 외에도 12,479자를 더 추가시킴으로써 모두 17,367자를 구현하였고, 조선 왕조 실록 CD-ROM에 수록된 기사의 개수는 총 362,161건이며, 그 속에 씌어진 문자의 수는 총 189,867,695이고, 입력된 문자의 수는 총 198,246,364자로 색인 생성을 위해 추출된 어절의 수는 41,083,164개이고, 2개 글자가 넘은 어절은 모두 1자 단위로 절단하여 색인을 생성한 결과, 색인 어휘의 총수는 74,122,921개가 생성되었다. 조선 왕조 실록 CD-ROM의 기능은 대체로 두 가지로 분류될 수 있다. 하나는 정보 열람 및 검색 기능이고, 또 다른 하나는 검색된 정보의 저장과 편집 기능으로 되어있다.

### 4. 전체 시스템 구조

본 연구에서는 제안하는 시스템은 크게 첫째, 문서간의 유사도 측정부 둘째, 클러스터링 측정부 그리고 셋째 Web 표현 부분 등으로 나뉜다




## Ⅲ. 유사도 측정 및 클러스터링

### 1. 색인 추출

색인어 작성의 특징은 처리하고자 하는 문서의 대리자 역할을 담당하고 있기 때문에 색인 생성시 색인 품질 여부를 잘 따져봐야 하고, 대상 자료가 가지고 있는 특징을 잘 표현 할 수 있는 중심어 파일(keyword file)색인을 자동적으로 생성할 수 있도록 해야 한다. 따라서 색인 작성의 목적은 주제에 의한 문서의 소재를 쉽게 파악할 수 있어야 하며, 주제의 영역을 정의함으로써 똥 다른 문서와 연계가능 할 수 있도록 한다. 또한 주어진 문서의 특정한 정보 요구에 적합 여부도 따져 봐야한다.

가) 형태에 따른 형태소 색인 추출 기법

나) 활용 정도에 따른 형태소 색인 추출 기법

다) HAM(HanGul Analysis Module)

본 연구의 대상이 되고 있는 역사적 기록문서 특징에 맞는 자동 색인어 생성 기능을 가진 도구가 필요하다. 이

를위해 자동 색인 단어 추출기로서는 HAM(HanGul Analysis Module)가 있다.[6]

### 2. 유사도 측정

문서 정보의 색인을 추출한 다음 키워드를 이용해 다음 중요한 작업이 유사도(similarity)일 것이다. 유사도를 측정하는 목적은 질의를 통해 표출된 정보 요구에 유사한 내용을 가진 문서들을 검색해 내는데 있다.[10] 이런 과정들을 통해 유사한 문서들을 한데 묶어줌으로써 좀 더 빠르게 검색할 수 있다. 유사도를 측정하는 여러 가지 방법이 있지만, 실제로 그 근간의 이루는 주요 개념들은 거의 비슷하다. 유사도의 측정은 유사 계수 공식을 사용하며 중요한 공식으로는 다이스 계수(Dice's Coefficient), 자카드계수(Jaccard's Coefficient), 코싸인계수(Cosine Coefficient), 중복도계수(Overlap Coefficient), 타니모트계수(Tanimoto Coefficient)가 있다.[11][12] 위에 제시한 공식을은 모두 비슷한 결과를 산출하므로 어느 공식을 선택할 것인가는 문제가 되지 않는다[12]. 그러나 본 연구에서는 HAM 방식에 의한 색인추출에서는 유사도 실험 추출 결과에 따라 다양한 유사 계수 값이 추출하는 것을 볼 수 있다.

다음은 각 유사계수 공식으로 |x| 와 |y|는 각 문서가 갖고 있는 색인어 수이고, |x∩y|는 공통되는 색인어 수, |x∪y|는 두 문서가 갖고 있는 다른 색인어의 합을 나타낸다. |x∩y|는 단순 일치계수로 각 계수는 이 값을 표준화시킨 것이다.[8]

유사도 측정은 대부분 Salton이 제시한 코사인 계수 방법을 이용하였지만, 본 연구에서는 문서간의 유사도 5가지 모든 경우 횟수를 테스트한 결과 중복도 계수에 의한 방식이 다른 어느 방식보다도 유사도가 많이 추하여 클러스팅에 적용하였다.

### 3. 클러스터링

클러스터링의 목적은 컴퓨터에 의해서 색인된 파일을 효율적으로 탐색하는데 있는 것이므로 이는 파일 조직이라고 볼 수 있다. 클러스터링 기법들은 문서에서 추출한 색인어나 또는 컴퓨터에 의해 추출된 키워드를 문서내용으로 질의를 위한 대리자로써 클러스터링을 형성한다. 문서들끼리 형성된 클러스터는 클러스터를 대표하는 클러스터 프로화일을 갖게 되며, 필요한 정보 요구에 가장 유사한 클러스터가 선택된다. 이는 밀접하게 상호 관련된 문서들은 동일한 정보요구에 대해 모두 적합하다는 것이다.[12][13]

대부분의 클러스터링 알고리즘을 생성하는데 고려해야 될 사항으로서는 클러스터링 형성 시 너무 많은 시간이 걸린다든가, 검색효율이나 검색시간 면에서 성능이 좋지 않는다든가, 형성된 클러스터의 속성[14]이 바람직하지 않는다든가 하는 근본적인 문제점을 안고 있다. 특히 공통적인 문제점으로는 이러한 클러스터링 알고리즘은 많은 양의 문서에는 효과적으로 사용하기 힘들다는 것이며, 실제로 대부분 수백 개 정도의 문서에만 실험적으로 적용되어 왔다.[12][15]

### 4 유사도 행렬에 의한 클러스터링

클러스터링을 접근 방법은 첫째, 문서간의 유사도를 측정하여 유사도 행렬을 작성하고 이로부터 계층적 클러스터를 형성하는 기법, 둘째, 직접 문서의 내용을 표현하는 색인어리스트를 가지고 클러스터를 형성하는 기법이 있다.

유사도행렬을 기초한 클러스터링은 n개 문서의 유사도 행렬을 작성하기 위해 n2 수준의 계산 작업이 필요한 반면, 재배치 클러스터링은 n 내지 n log n 수준의 계산이 필요하다.[16] 따라서 유사도 행렬에 의한 클러스터링은 파일의 크기가 큰 경우에는 너무 많은 계산작업을 요하므로 적합하지 않다. 반면에 재배치 클러스터링은 클러스터링 시간은 빠르나 대부분 검색 효율 면에서 떨어지거나, 또는 문서의 입력순서에 따라 클러스터링 결과가 변화하는 문제점을 갖는다. 자기발견적 클러스터링은 재배치 클러스터링라고도 한다. 이는 클러스터의 센트로이드와 문서간의 유사도 측정에 기초하므로 클러스터 센트로이드의 형성이 선행되어야 한다. 본 연구에서는 센트로이드 값을 질의어로 대치하여 처리하는데 여기서 처리란 문서와 클러스터 센트로이드(질의어)와의 유사도 측정작업을 의미한다.

## Ⅳ. 실험 및 평가

### 1. 실험 방법 및 자료 구축

문서간 유사도 측정 자료는 역사적 기록문서의 대표인

조선 왕조 실록 CD-ROM을 기본 자료를 이용했다. 조선 왕조 실록 CD-ROM에 기록되어있는 자료에서 KS 표준한자인 4,888자 외에도 12,479자를 더 추가시킴으로써 모두 17,367자를 구현하였고, 조선 왕조 실록 CD-ROM에 수록된 기사의 개수는 총 362,161건이며, 그 속에 씌어진 문자의 수는 총 189,867,695이고, 입력된 문자의 수는 총 198,246,364자로 색인 생성을 위해 추출된 어절의 수는 41,083,164개이고, 2개 글자가 넘은 어절은 모두 1자 단위로 절단하여 색인을 생성한 결과, 색인 어휘의 총수는 74,122,921개가 생성되었다. 실험을 통해 문서 유형의 차이가 유사도, 클러스터링의 성능 평가에 영향을 미칠 수 있기때문에, 본 연구에서는 요구 사항 문서들을 효율적으로 분류하고 관리하기 위하여 제3장에서 제시한 역사적 기록문서 특징에 맞는 자동 색인어 생성 기능을 가진 HAM(Hangul Analysis Module)를 사용, 색인어 행렬을 만들어 유사도 측정 기법에 사용하였다. 조선왕조실록에 기록된 모든 자료를 처리하기에는 너무 많은 자료이기 때문에, 제안된 문서간 유사도 측정 기법의 효율성 평과를 위하여 일부분을 추출해서 실험 데이터를 사용하였다.

먼저, 인조 실록 2년(1624년)차 1월부터 12월까지의 자료를 실험데이터로 사용하였다. 모두 523개의 문서를 가지고 한국어 자동 색인기 HAM의 색인결과 5079개의 색인이 생성되었으며, 문서 당 최소 색인 개수 1개부터 시작해서 문서 당 최대 색인 개수 55개까지 추출되었으며, 한 문서 당 색인 개수는 평균 16개의 색인이 추출되었다.

문서간의 유사도 측정 기법은 문서와 색인어의 관계를 그림과 같이 벡터로 표현하여 문서-색인어 행렬을 작성하고, 문서와 문서간의 유사도를 측정하며 그림과 같이 문서-문서 유사 계수 행렬을 형성한다.

## 2. 실험 결과

### 2.1 유사도 측정

유사도 측정은 제3장에서 제시한 여러 가지가 있고, 이런 공식들은 비슷한 결과를 산출하므로 어느 공식을 선택할 것인가는 문제가 되지 않는다. 대부분 코싸인 계수 방식과 다이스 계수 방식이 비교적 널리 사용되고 있다.[8] 그러나 역사적 기록문서에서 추출한 색인어를 이용하여 다섯가지의 유사도를 측정한 결과 많은 차이를 보이고 있다. 따라서 본 연구에서는 유사도 측정 방식에서 유사도 계수가 많이 높은 방식이 중복도 방식에서 추출된 자료를 선택하였다.

| 유사계수값 | 0.5이상 | 0.6이상 |
|---|---|---|
| 다이스 계수 방식 | 9 | 4 |
| | 5% | 4% |
| 자카드 계수 방식 | 0 | 0 |
| | 0% | 0% |
| 코싸인 계수 방식 | 15 | 7 |
| | 14% | 8% |
| 중복도 계수 방식 | 82 | 73 |
| | 77% | 86% |
| 타니모토 계수 방식 | 0 | 0 |
| | 0% | 0% |
| 소 계 | 106 | 84 |
| 합 계 | 324 | 324 |

인조 실록 2년(1624년)차 1월부터 12월까지의 자료를 실험데이터로 사용하여 모두 523개의 문서를 가지고 한국어 자동 색인기 HAM의 색인결과 5079개의 색인이 생성되었으며, 추출된 색인수를 가지고 유사값을 계산한 결과, 표[ ]와 같은 수치가 나왔다. 유사계수 0.1부터 0.9까지 각 단계별로 유사도를 추출한 결과 비교적 널리 사용되고 있는 코싸인 계수와 다이스 계수 방식 보다는 중복도에 의한 계수가 상대적으로 높은 값이 나와 조선왕조 실록 색인어를 이용한 유사도 계수 방식에는 중복도 계수 방식이 적절하다는 것을 알 수 있다. 그 원인을 파악하자면, 첫째로 역사적 기록 문서를 기술하는데 있어서 왕의 출생, 성격, 일화(逸話), 즉위에 관련된 내용 등을 포함하는 각 실록 머리의 일진(日辰) 표시 없이 작성된 기사로 총서를 했기 때문이다. 둘째로 역사적 기록문서의 형식이 대화(對話), 운문(韻文), 전교(傳敎), 교서(敎書), 상소(上訴) 등 일정한 형식으로 기술했기 때문에 서로 다른 내용이라도 형식이 비슷하다. 셋째로 인명(人名), 관직명(官職名), 관서명(官署名), 작호(爵號), 제도 용어 등의 고유 명사들이 자주 표현되었다. 넷째로 국사 편찬위원회가 "조선왕조실록분류사편찬요강"에 의거 항목 분류를 기본항목 40개, 세부항목 161개로만 사용했기 때문에 중복이 많을 수밖에 없다.

#### 4.2.2 클러스터링

클러스터링방법은 두 가지 분류하고 있다. 첫째는 문서간의 유사도를 측정하여 유사도 행렬을 작성하고 이를

계층적 클러스터링을 형성하는 방법과 둘째로 문서를 직접 표현하는 색인어 리스트를 가지고 클러스터를 형성하는 방법이다. 본 연구에서는 두 번째 방법인 자기 발견적 클러스터링 기법중에 다돌라식 기법으로 접근한다.

다돌라식 알고리즘의 특성은 문서의 입력순서에 따라 클러스터의 내용이 달라지므로 고도의 순서 의존적이라는 점과 클러스터 중복성이 허용된다는 점이다[17]

### 3. 평가

클러스터링을 위해 324건의 문서 중에서 유사도 계수가 0.6 이상인 60건으로 4개의 그룹으로 분류해서 같은 클러스터링 그룹으로 찾기 위해 10회 반복을 하였다. 최종결과를 확인한 결과 4번째까지는 활발한 그룹으로 이동하고, 5번째부터는 10번째까지는 변동이 없었다. 전혀 관련이 없는 문서는 group 0로 이용되었고, 실제로 자료를 확인 해본 결과 group 2는 질의 문서와 가장 가까운 그룹에 속한 문서들이 나타났다.

## V. 결 론

본 연구는 역사적 기록 문서에서 이벤트들의 효율적인 검색에 있어서 축적된 자료를 특정한 정보 요구 시, 요구에 부합하는 문서를 어떻게 찾아내느냐에 대해서 유사도 측정 및 그 와 연계하여 클러스터링을 효과적으로 할 수 있다는 것을 보여 주었다.

먼저 역사적 기록 문서인 조선왕조실록 CD-ROM에 기록된 자료를 각 일진 별로 구분하여 일진별 문서를 만들고, 문서를 대표할 수 있는 HAM에 의해 색인어를 만들었다. 그러나 문서의 특성상 추출된 색인어는 다시 정리하여 일진 문장을 대표할 수 있는 색인어로 재구성하였다. 그 재구성한 이유는 제4장 실험에서 나타났듯이 역사적 기록 문서를 작성할 때 일정한 형식으로 자료를 표현했기 때문에 서로 다른 내용이라도 인명, 관직명, 관서명, 작호, 제도 용어 등의 고유 명사들이 자주 표현되었기 때문이다.

본 연구에서 문서간의 유사도 방법을 찾아내기 위해 여러 가지 방법을 실험한 결과 중복도 계수 방식이 가장 적절하다는 것을 발견했으며, 중복도 계수에 채택된 문서를 이용해서 클러스터링 실험을 하였다.

클러스터링은 자기 발견적 클러스터링에서 클러스터의 센트로이드와 문서간의 유사도 측정에 기초하므로 클러스터 센트로이드 형성을 질의 문서로 채택하였다. 즉 질의 문서를 센트로이드 문서로 간주하고 클러스터링 실험을 하였고, 실제 작업한 결과 질의 문서와 클러스터링 그룹핑한 것이 적절하게 표현되었다.

본 논문에서 좀 더 보완해야 할 것은 방대한 조선 왕조 실록을 모두 실험 대상으로 하였다면, 더 정확한 클러스터링 그룹핑을 했을 것이다. 따라서 조선 왕조 실록 CD-ROM 모든 자료를 색인어, 유사도, 클러스터링 실험에 이용하는 큰 작업을 해야할 것이다. 그리고 선택된 문서들을 국제 표준 멀티미디어 문서 XML로 표현할 수 있도록 연구를 계속해야 할 것이다.

## 참고문헌


[1] 정보 저장 및 검색 , Robert R. Korfhage 원저, 류근호 외 1인 옮김, 시그마프래스, 2000]

[2] 조선 실록, 재단 법인 민족문화추진 위원회, 1987년

[3] 이조 실록, (평양) 사회과학원 민족고전연구소, 남한 여강 출판사, 1993년

[4] 두산 세계대백과 EnCyber

[5] CD-ROM 국역 조선왕조실록 안내서

[6] http://www.nip.kookmin.ac.kr/HAM/kor

[7] 조선 왕조 실록 CD-ROM

[8] Salton G. and McGill M., J., Introduction to Modern Information Retrieval(Computer Series), New York : McGraw-Hill, 1983

[9] Ash R., Information Theory, New York : Wiley Interscience, 1965

[10] 김학수 외 3, "유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현" 정보과학회 제27권 제 1호(2000.1)

[11] 자카드 계수와 타니모토 계수 공식의 분모는 모

두 두 문서 속세 출현한 다른 용어의 수를 산출하는 것으로 같은 것이며 따라서 계수값도 같아진다.

[12] 정영미, "정보검색론", 구미무역(주) 출판부, 1993. p182

[13] C.J. van Rijsbergen, Information Tetrieval(2nd ed.) (London: Butterworths, 1979), p45

[14] 문서 입력 순서에의 독립성, 문서양의 증가나 착오발생으로부터의 안정성 등이 클러스터를 평가하는 대표적 속성이다

[15] 크랜필드 실험장서인 1400개의 문서를 싱글링크 방식으로 클러스터링한 실험과[C. J. van Rijbergen and W. B. Croft, "Document Clustering:An Evaluation of Some Experiments with the Cranfield 1400 Collection," IPM 11: 171~182; 1975] FIRST 시스템에서 2만개의 문서를 다톨라의 알고리즘을 사용하여 클러스터링한 것은 예외적인 경우라 볼 수 있다.[R.T. Dattola, "FIRST : Flexible Information Retrieval System for Text," JASIS 30(1):9~14 : 1979].

[16] P. Willet, "Document Clustering Using an Inverted File Approach," JIS 2 : 223!231 : 1980

[17] Dattola, R. T.," A Fast Algorithm for Automatic Classification," JLA 2(1):31-48 : 1969

## 저 자 소 개

한 광 덕

1988년 한국외국어대학 컴퓨터 공학과 석사

1999년 한국외국어대학 컴퓨터 공학과 박사수료

관심분야:멀티미디어

1997- 현재 상지영서대학 컴퓨터정보기술과 교수