# 고전 문서의 효과적인 검색 결과 하이라이팅

## Effective Highlighting Retrieval Results of Historical Documents

정창후, 최윤수, 김광영, 서정현, 윤화묵
한국과학기술정보연구원

Jeong Chang-Hoo, Choi Yun-Soo, Kim Kwang-Young,
Seo Jeong-Hyeon, Yoon Hwa-Mook
Korea Institute of Science and Technology Information

**요약**

본 논문에서는 고전 문서가 XML 형태로 전산화된 이후에, 의미적 특징을 최대한 손상시키지 않고 검색 결과를 효과적으로 하이라이팅하는 방법에 대해서 설명한다. 특히, 고전 문서의 특성을 최대한 고려하여 하이라이팅 문자열 비교를 수행하였다. 또한, XML 문서의 특성을 고려하여 하이라이팅 태그 삽입 시에 다양한 처리를 수행하였다.

**Abstract**

In this paper, we introduce a method to effectively highlight retrieval results without impairing meaningful features after historical documents were digitized into XML format. Especially, making the best of the features of historical documents, we perform string matching for the highlighting. Also, considering the features of the XML document, we carry out various processes when highlighting tag is inserted.

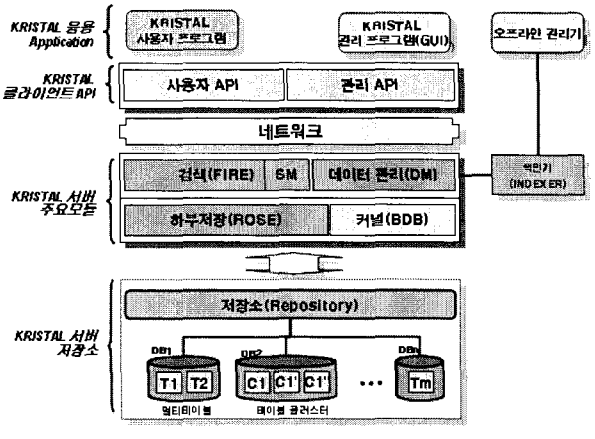
## I. 서 론

고전 문서는 우리나라의 역사와 문화, 선인들의 사상, 사회 풍속과 제도 등을 후손들에게 전달해 줌으로써 연구와 교육을 위한 가치 있는 정보 자원의 역할을 한다. 고전 문서가 정보 자원으로서의 역할을 다하기 위해서는 접근과 이용이 용이해야 함에도 불구하고 보존상의 이유로 많은 제약이 있어 왔다. 역사적 가치와 보존 가치가 높은 고전 문서의 훼손을 방지하고 접근을 용이하게 하기 위해서는 고전 문서의 전산화가 필요한데, 최근의 경제 성장 및 일반인의 역사적 정체성에 대한 욕구가 증가함에 따라 대규모의 고전 문서 전산화 작업이 활발히 진행되고 있다. 고전 문서 전산화 작업에서 가장 어렵고 비용이 많이 소요되는 분야는 고전 문서의 의미적 특징을 최대한 손상시키지 않으면서 서비스를 구축하는 일이다[1]. 본 논문에서는 고전 문서 서비스 기능의 하나인 효과적인 검색 결과 하이라이팅 방법에 대해서 설명하도록 한다.

## II. KRISTAL-IRMS에서 고전 문서의 하이라이팅

### 1. KRISTAL-IRMS

정보검색엔진(IRS)의 모든 기능과 데이터베이스관리시스템(DBMS)의 일부 기능을 밀결합한 정보검색관리시스템 KRISTAL-IRMS(Knowledge Retrieval In Science & Technology Affiliated Literatures - Information Retrieval Management System)[2]는 문헌 정보 데이터와 XML 구조 문서를 하나의 데이터베이스에서 저장, 관리하고 검색할 수 있도록 설계 및 개발되었다.




▶▶ 그림 1. KRISTAL-IRMS의 전체 구조

그림 1은 KRISTAL-IRMS의 전체 구조이다. 문서의 검색과 관리를 동시에 지원하기 때문에 다양한 모듈이 존재하나, 본 논문에서는 XML로 작성된 고전 문서를 효과적으로 하이라이팅하는 모듈에 대해서 설명하도록 한다. 클라이언트-서버 구조로 되어있는 KRISTAL-IRMS에서 하이라이팅은 사용자 질의어를 알고 있는 클라이언트 쪽에서 기능 구현을 할 수도 있으나, 다음과 같은 2가지 이유로 인해서 서버 내부 기능으로 구현한다.

첫째, 사용자가 입력한 질의어는 서버 내에서 형태소 분석 및 질의어 확장 과정을 거치면서 시스템 질의어로 확장이 되는데, 이렇게 확장된 시스템 질의어를 사용자가 알기 위해서는 KRISTAL-IRMS 클라이언트 API[3]를 호출해서 정보를 가져가야 한다. 정보를 가져간 후에 검색 필드별 질의어를 이용하여 하이라이팅을 처리하면 되는데, 애플리케이션 개발자에게 불편한 작업이 될 수 있다.

둘째, 한글 질의어를 입력해서 한자로 작성된 문서를 하이라이팅하거나 혹은 이체자나 두음법칙을 적용하여 하이라이팅해야 하는 경우가 있는데, 이러한 작업을 클라이언트 쪽에서 수행하기 위해서는 많은 언어 자원이 요구된다. 클라이언트에서 이러한 자원을 유지하는 것은 매우 비효율적이기 때문에, KRISTAL-IRMS에서는 문서 하이라이팅을 서버의 기능으로 제공한다.

## 2. 고전 문서 하이라이팅의 특성

일반적으로 고전 문서는 국가적인 프로젝트로 수행되어 제작되는데, 고전 문서의 대부분은 한자로 이루어져 있어서 문서를 디지털화 하는데 전문 인력이 필요하다. 이렇게 제작되어진 고전 문서는 그 문서 하나하나가 커다란 의미를 갖고 있기 때문에, 검색 시 한 문서라도 누락시키지 않기 위해서 한자 색인을 효과적으로 지원해야 한다. 마찬가지로 이러한 색인 과정을 거쳐서 검색된 문서에 해당 질의어를 색인과 같은 원리로 하이라이팅해주는 메커니즘이 필요하다. 고전 문서에는 다음과 같은 특성이 존재한다.

첫째, 고전 문서는 공백의 구분이 없이 한자로 기록된다. (예 : "又如李慶祿·李舜臣者欲用之, 竝參酌議啓。")

둘째, 고전 문서를 구성하는 한자에는 이체자가 존재한다. (예 : 劍 = "劍", "劍", "劒", "劍", "劒", "劒", "釼", "釼", "鎩")

셋째, 고전 문서를 서비스 받는 대다수의 사용자는 한국어 발음으로 한자를 검색하기를 원한다. (예 : 충무공 = "忠武公")

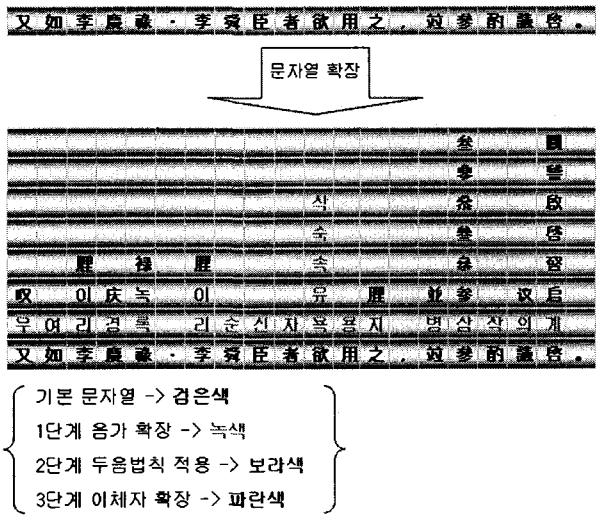
넷째, 고전 문서를 구성하는 한자에는 1개 이상의 한국어 음가가 존재한다. 따라서 동형 이음어를 처리해줘야 한다. (예 : 樂 = "악", "락", "요")

다섯째, 고전 문서를 구성하는 한자에 한국어 음가를 적용할 때 두음법칙을 함께 고려해줘야 한다. (예 : 樂 = "락", "낙")

여섯째, 새롭게 전산화되고 있는 대다수의 고전 문서는 번역되어 한문-국역이 혼재한다. 따라서 한글 데이터와 한자 데이터를 모두 처리할 수 있어야 한다.

이와 같이 고전 문서는 전산화되어서 존재하는 문서의 기본 문자뿐만 아니라 그 문자를 편리하게 해석하고 활용하기 위한 다양한 의미론적 처리를 수행해줘야 한다. 예를 들어, 고전 문서에서 한글을 이용하여 검색하였을 경우에 한자의 어느 부분이 검색되었는지를 알기가 쉽지 않기 때문에 해당 질의어를 하이라이팅해줄 필요가 있다.

하이라이팅 문자열 탐색을 위해서 기본 문자열이 확장되는 과정은 그림 2와 같다.




▶▶ 그림 2. 문자열 확장 과정

처음에는 기본적으로 한자로 작성된 문자열이 존재한다. 여기에 음가 확장 및 동형 이음어 처리를 수행하고, 다음으로 두음 법칙을 적용하여 단순 문자열을 확장한다. 끝으로 이체자 확장을 수행하여 동일한 뜻을 지닌 여러 형태의 한자를 같은 문자로 해석한다. 결과적으로 기본 문자열 대신에 확장 문자열을 사용함으로써 고전 문서의 하이라이팅을 효과적으로 지원할 수 있다.

## 3. XML 문서 하이라이팅의 특성

고전 문서를 전산화하는 과정에서 사용이 간편하고 재사용성 및 확장성이 뛰어난 XML 문서가 주로 사용되고 있다. XML(eXtensible Markup Language)[4] 문서는 구조적 데이터 표현 및 문서 교환의 표준으로 1998년 W3C에서 채택된 후 전자상거래, 전자책 등 많은 분야에서 활용되고 있다. XML 문서를 하이라이팅하기 위해서는 다음과 같은 XML 문서의 특성이 하이라이팅 태그 삽입 시에 반영되어야 한다.

첫째, XML 문법에 어긋나지 않아야 한다. XML 문서 원본에는 "<성>이</성><이름>순신</이름>"와 같이 되어 있더라도 스타일시트를 적용한 사용자 브라우저 상에는 "이순신"이라고 표시된다. 따라서 위의 문서를 브라우저에서 바라보는 최종 사용자는 "이순신"으로 검색을 하였을 경우에 "이순신"이 하이라이팅되기를 기대한다. 따라서 위의 예제를 하이라이팅하였을 경우에는 "<성><Highlight color='red'>이</성><이름>순신</Highlight></이름>"이 되지만 XML 문

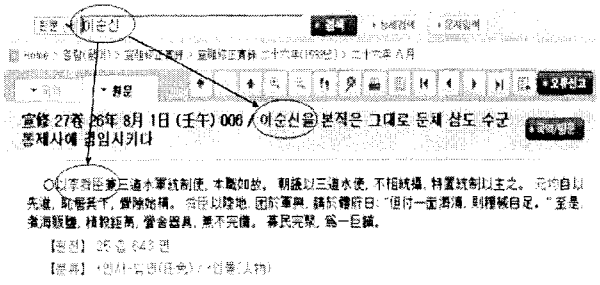
법에 맞춰주기 위해서 "<성><Highlight color='red'>이</Highlight></성><이름><Highlight color='red'>순신</Highlight></이름>"와 같은 결과를 생성해 낸다. 물론 XML 문서의 "Highlight"라는 태그의 텍스트를 "color"라는 속성으로 지정된 색깔로 표시하도록 미리 스타일시트를 구성해 놓아야 한다. 이러한 하이라이팅 태그는 파라메터에 의해서 조절이 가능하다.

둘째, 질의어를 XML 문서에서 탐색해 나가다가 중간에 엘리먼트가 발견되면 엘리먼트의 이름, 속성 이름, 속성 값과 같은 부가적인 정보는 문자열 비교 대상에서 제외하여야 한다. "성웅이순신"으로 검색을 하였을 경우에 "<성>이</성><이름>순신</이름>" 문자열에서 "성"이나 "이름"과 같은 엘리먼트 이름은 문자열 탐색에서 비교를 수행하지 않기 때문에 불필요한 비교 횟수를 효과적으로 줄일 수 있다.

셋째, 하이라이팅 태그의 중복을 제거하여 포괄적으로 만들어야 한다. 사용자 질의어로 "정보&검색&시스템"으로 검색을 하였을 경우에 "정보", "검색", 그리고 "시스템"의 3가지 단어를 가지고 하이라이팅을 수행하게 된다. 문서에 존재하는 각 단어들이 하이라이팅되는 것은 당연할 뿐만 아니라, "정보검색시스템"이라는 단어가 존재할 경우 각각의 단어에 대해서 하이라이팅을 하게 되면 "<Highlight color='red'>정보</Highlight><Highlight color='red'>검색</Highlight><Highlight color='red'>시스템</Highlight>"이 생성되는 것이 아니라 "<Highlight color='red'>정보검색시스템</Highlight>"이 생성된다. 이와 같이 최장 일치 방식을 사용함으로써 불필요하게 생성되는 정보를 줄일 수 있을 뿐만 아니라 문서의 간결함도 함께 유지할 수 있다.

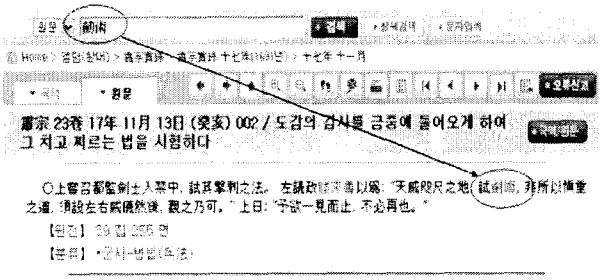
## III. 실 험

실험은 조선왕조실록 데이터를 이용하여 수행하였다. 조선왕조실록(朝鮮王朝實錄)은 조선시대 역대 임금들의 실록을 합쳐서 부르는 책 이름이다. 즉 태조강헌대왕실록(太祖康獻大王實錄)으로부터 철종대왕실록(哲宗大王實錄)에 이르기까지 472년간에 걸친 25대 임금들의 실록 28종을 통틀어 지칭하는 것이다. 조선왕조실록은 특정한 시기에 특정한 사람들이 의도적으로 기획하여 편찬한 역사서가 아니라, 역대 조정에서 국왕이 교체될 때마다 편찬한 것이 축적되어 이루어진 것이다. 조선왕조실록 홈페이지[5]는 이러한 조선왕조실록을 XML 형태로 전산화하여 KRISTAL-IRMS를 기반으로 서비스하고 있기 때문에 고전 문서의 특성과 XML 문서의 특성을 모두 처리해 줄 수 있다.



▶▶ 그림 3. 음가 확장 및 두음 법칙 적용 하이라이팅

그림 3은 음가 확장 및 두음 법칙 적용 하이라이팅의 예를 보여준다. 질의어로 "이순신"을 입력하였지만 문자열의 음가 확장 및 두음 법칙 적용을 통하여 "李舜臣"을 효과적으로 하이라이팅하는 것을 살펴볼 수 있다.



▶▶ 그림 4. 이체자 확장 하이라이팅

그림 4는 이체자 확장 하이라이팅의 예를 보여준다. 질의어로 "劒術"을 입력하였지만 문자열의 이체자 확장을 통하여 "劍術"을 효과적으로 하이라이팅하는 것을 살펴볼 수 있다.

## IV. 결론 및 향후 연구

본 논문에서는 고전 문서가 XML 형태로 전산화된 이후에, 의미적 특징을 최대한 손상시키지 않고 검색 결과를 효과적으로 하이라이팅하는 방법에 대해서 살펴보았다. 특히 음가 확장, 동형 이음어 처리, 두음 법칙 적용, 이체자 확장 등과 같은 고전 문서의 특성을 최대한 고려하여 하이라이팅 문자열 비교를 수행하였다. 또한, XML 문서의 특성을 고려하여 하이라이팅 태그 삽입 시에 다양한 처리를 수행하였다.

향후 연구로는 사용자의 다양하고 복잡한 요구 사항을 충분히 반영할 수 있도록 하이라이팅 기능에 대한 보다 상세하고 다양한 명세 작업이 필요하다. 적용하는 분야에 따라서 사용자의 의견이 상이할 수 있기 때문에 하이라이팅 방법에 좀 더 융통성을 부여할 필요가 있다.

## ▌참 고 문 헌▐

[1] 최윤수, 서정현, 진두석, 정창후, "효율적인 고전문서 관리 및 검색을 위한 정보검색 관리 시스템의 구현", 한국인터넷정보학회 춘계학술발표논문집, 제6권, 제1호, pp.369-372, 2005.
[2] 정보검색관리시스템 KRISTAL-IRMS, "http://www.kristalinfo.com"
[3] 주원균, 정창후, 이민호, 양명석, 최윤수, 최기석, "KRISTAL-2002를 위한 JAVA 사용자 API의 설계 및 구현", 한국정보과학회 제 31회 추계학술발표논문집, 2004.
[4] Extensible Markup Language(XML) 1.0, "http://www.w3.org/XML"
[5] 조선왕조실록 홈페이지, "http://sillok.history.go.kr"