공공사이트 게시판 자료의 기록관리를 위한 자동 분류 시스템

Automatic classification system for record management of bulletin board on public website

남은경, 연세대학교 문헌정보학과, namek@yonsei.ac.kr
안혜림, 연세대학교 문헌정보학과, hrahn@yonsei.ac.kr
송민, 연세대학교 문헌정보학과, min.song@yonsei.ac.kr

Eunkyung Nam, Dept. of LIS, Yonsei University
Hye-Rim Ahn, Dept. of LIS, Yonsei University
Min Song, Dept. of LIS, Yonsei University

웹의 발달과 전자정부의 지향으로, 행정기관의 웹사이트를 통한 민원처리가 증가하고 있다. 게시판을 통해 이용자가 민원을 제기하면, 각 기관에서는 담당자를 배정해 처리하지만 해당 게시물을 공기록으로 보존하지는 않는다. 공공사이트를 통한 투명한 행정을 위해서는 게시물도 공기록물로 보존하는 체계가 마련될 필요가 있다. 이를 위해, 정부기능연계모델(BRM)을 기준으로, 공공사이트의 게시글을 자동으로 분류하는 시스템을 구현하였다.

1. 서 론

웹의 발달과 전자정부의 지향으로, 각 행정기관의 웹사이트를 통해 다양한 행정 업무가 처리되고 있다. 그 중, 게시판은 이용자들이 해당 기관과 소통하는 창구로 활발하게 이용되고 있는데, 이용자가 게시판을 통해 민원을 제기하면 각 기관에서는 담당자를 배정해 답변하도록 되어 있어, 이용자는 이를 해당 기관의 공식적인 입장으로 받아들이게 된다.

이처럼 웹사이트를 통해 제기되는 민원은 해당 기관의 공식 행정 절차에 따라 처리되지만, 그 게시물이 공기록으로 보존되고 있지는 않다. 정책 오류를 보여주는 게시물을 은폐하려 한 미국 Arlington Country의 사례(Barry, 2004)에서와 같이 공공게시판의 글이 정부의 정책이나 의도에 반하는 경우 은폐 또는 삭제될 수 때문에, 투명한 행정을 위해서는 공공사이트의 게시판 자료도 공기록물 관리 체계에 편입되어 보존·관리되어야 한다.

하지만, 게시글의 공기록화는 공무원들의 업무 하중을 증가시키는 원인이 될 수 있다. 업무 분장 시, 일반 직원에게 기록 관리 업무를 맡기지 않는 것이 업무의 효율성에 긍정적 영향을 미친다(Sprehe & McClure, 2005). 실제 경기도 정부를 비롯한 행정기관은 기록물 관리 시스템을 도입하는 등 체계적인 관리를 위한 노력을 하고 있으나 기록물 업무를 전담하는 전문 인력은 부족한 실정이고, 해당 업무는 관련 정보를 저장하고 관리하는 수준에 그치고 있다(한성산, 2010). 때문에, 공공게시판 자료의 공기록화에 대한 논의는 활발하게 진행되지 않고 있다.

이에 따라, 본 연구는 공공게시판의 게시물을 자동적으로 분류해주는 시스템을 구현하고자 한다. 시스템은 crawling, 한글 형태소 분

석, 정부기능연계모델(Business Reference Model, BRM)에 따른, XML 형식의 결과 파일 생성 과정을 한 번에 처리한다. 시스템은 Java를 이용해 구현하였다. 자동 분류된 결과는 적합성 평가를 수행하였다.

2. 시스템 구현

2.1 Crawling

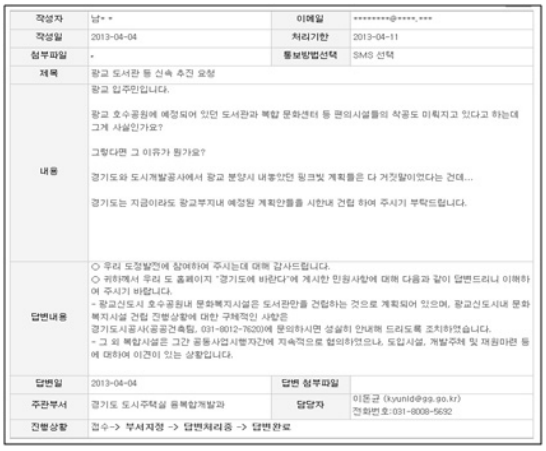
Crawling 대상 게시판은 경기도 공식 홈페이지의 '경기도에 바란다' 게시판이다. 경기도는 웹을 이용한 다양한 행정 서비스를 제공하고 있고, 42,000여 건(2013.08 현재)의 게시물이 축적되어 있어 분석 대상으로 적절하다고 판단하였다. 해당 게시판은 이용자가 민원을 제기하면 게시판 관리자가 주관 부서에 해당 민원을 접수하고, 주관 부서의 담당자가 답변을 하도록 되어 있다. 이용자는 이를 경기도청의 공식적인 입장이라 받아들이게 되므로, 법적인 의무는 없으나 사실상의 민원 처리가 이루어지고 있다고 할 수 있다.

게시글 항목 중 작성자, 이메일, 작성일, 처리기한, 제목, 내용, 답변내용, 답변일, 주관부서, 담당자의 10개 필드를 추출하였다(<그림 1> 참조). 민원 처리가 이뤄진(즉 답변이 달린) 글에 대해서만 crawling을 실시하였다. crawling에는 Jsoup crawler를 사용하였다.

2.2 형태소 분석

crawling한 데이터 중 게시판 글의 제목, 내용, 답변 내용만을 대상으로 형태소 분석을 실시하였다. Lucene 한국어 형태소 분석기의 MorphAnalyzer를 이용하였고, 분석 결과에서 명사만을 추출하였다.

2.3 분류



작성자	남* *	이메일	********@****.***
작성일	2013-04-04	처리기한	2013-04-11
첨부파일	*	통보방법선택	SMS 선택
제목	광교 도서관 등 신속 추진 요청		
내용	광교 입주민입니다. 광교 호수공원에 예정되어 있던 도서관과 복합 문화센터 등 편의시설들의 착공도 미뤄지고 있다고 하는데 그게 사실인가요? 그렇다면 그 이유가 뭔가요? 경기도와 도시개발공사에서 광교 분양시 내놓았던 밍크빛 계획들은 다 거짓말이었다는 건데... 경기도는 지금이라도 광교부지내 예정된 계획안들을 시한내 건립 하여 주시기 부탁드립니다.		
답변내용	○ 우리 도정발전에 참여하여 주시는데 대해 감사드립니다. ○ 귀하께서 우리 도 홈페이지 "경기도에 바란다"에 게시한 민원사항에 대해 다음과 같이 답변드리니 이해하여 주시기 바랍니다. - 광교신도시 호수공원내 문화복지시설은 도서관만을 건립하는 것으로 계획되어 있으며, 광교신도시내 문화복지시설 건립 진행상황에 대한 구체적인 사항은 경기도시공사(공공건축팀, 031-8012-7620)에 문의하시면 성실히 안내해 드리도록 조치하였습니다. - 그 외 복합시설은 그간 공동사업시행자간에 지속적으로 협의하였으나, 도입시설, 개발주체 및 재원마련 등에 대하여 이견이 있는 상황입니다.		
답변일	2013-04-04	답변 첨부파일	
주관부서	경기도 도시주택실 융복합개발과	답당자	이동균 (kyunld@gg.go.kr) 전화번호:031-8008-5692
진행상황	접수-> 부서지정 -> 답변처리중 -> 답변완료		

<그림 1> 게시글 화면

1) 경기도 BRM

본 연구에서는 BRM을 기준으로 게시판 글을 분류하였다. BRM은(<그림 2> 참조) 정부기능을 범정부 차원에서 업무 및 서비스 중심으로 분류하고, 다양한 정보를 기능과 연계하여 종합적으로 관리함으로써, 공통 업무기반을 제공하는 업무참조모델(정부혁신지방분권위원회, 2003)이다.

<그림 2> 정부기능연계모델(BRM)

본 연구는 경기도 정부의 공식 웹페이지 게시판 글을 대상으로 하였으므로, 정보공개청구를 통해 경기도에서 사용하는 BRM(이하 경기도 BRM)을 엑셀 파일로 확보하였다.

경기도 BRM은 15개 정책분야, 48개 정책

영역, 129개 대기능, 420개 중기능, 1,374개 소기능, 16,557개 단위과제로 구성되어 있었다. 이중 중복 명기된 단위과제를 삭제하여 총 7,022개 단위과제를 분류에 적용하였다.

경기도 BRM에는 띄어쓰기가 되어 있지 않아, 단어 단위 분석을 위해 공백을 삽입하였다. 띄어쓰기 작업 후, 변별력이 없는 단어(및, 등, 관련, 관한, 대한, 기타, 중, 일반, …)들은 일괄적으로 제거하였다. 전처리 후의 BRM은 <그림 3>과 같다.

기능별 분류체계

기관명 : 경기도

분류체계

정책분야	정책영역	대기능	중기능	소기능	단위과제
공공질서 안전	경찰	수사	특별 사법 경찰 관리 작무	특사경 계획 수립	특별 사법 경찰 관리
공공질서 안전	경찰	수사	특별 사법 경찰 관리 작무	특사경 직무 수행	특사경 직무 수행
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구급 보고
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구급 업무
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구급
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구급 증명
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구조 구급 활보, 분기보
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구급 업무	구조 업무
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 법령 정비	구급 대편성 과 운영
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 법령 정비	구조 구급법령 제정 정비
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	119 유실물 지키미 제도 운영
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	119 안명구조 유공대상
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	U119 서비스 업무
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급 대책 협의회 계획 운영
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급 업무 운영 계획
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급 활보
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급 활보반기 보고
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구급 활동 사례 수집 홍보
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구조메뉴얼 작성
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구조 교육 포상
공공질서 안전	재난방재 민방위	구조 구급	구급 업무 정책	구조 구급 종합 계획	구조 구급 업무 운영 계획 수립 운영

<그림 3> 데이터 전처리 후 경기도 BRM

데이터 전처리 후 탭으로 분리된 텍스트 파일로 저장, 본래의 6단계 구조를 유지하였다. Java 프로그램에서 텍스트 파일을 행 단위로 loading하여 단위과제별로 처리하였다.

2) 경기도 BRM에 따른 분류

형태소 분석 결과 텍스트에서 각 단위과제에 포함되어 있는 단어의 출현 빈도를 계산하고, 출현 빈도에 단계별 가중치를 적용하여 점수를 산출하였다. 점수가 가장 높은 단위과제에 해당 게시글이 분류되도록 하였다.

BRM 단계별 가중치는 각 BRM 단계가 가진 변별력, 즉 단계별 항목수를 기준으로 하였다. 경기도 BRM에 따라 각 단계별 가중치는 <표1>과 같다.

<표 1> 경기도 BRM의 단계별 가중치

분류체계	가중치
정책분야	15 / 7022 = 0.0021
정책영역	48 / 7022 = 0.0068
대기능	129 / 7022 = 0.0174
중기능	421 / 7022 = 0.0598
소기능	387 / 7022 = 0.1957
단위과제	7022 / 7022 = 1

위의 가중치를 적용한 전체 점수는 score = $\sum$ $count_i$(빈도수) * $weight_i$(가중치)로 구한다.

2.4 XML 파일 출력

게시글에서 crawling한 10개 필드와 BRM에 따른 분류 6개 필드를 보존을 위해 XML 파일로 출력하였다. 해당 파일은 (게시글 번호).xml로 저장하였고, 출력된 XML 파일은 <그림 4>와 같다.

```
<field id="10">
<data key="label">정책분야</data><data key="content"> 공공행정</data>
</field>
<field id="11">
<data key="label">정책영역</data><data key="content"> 행정</data>
</field>
<field id="12">
<data key="label">대기능</data><data key="content"> 행정 지원 </data>
</field>
<field id="13">
<data key="label">중기능</data><data key="content"> 기획</data>
</field>
<field id="14">
<data key="label">소기능</data><data key="content"> 종합 기획조정</data>
</field>
<field id="15">
<data key="label">단위과제</data><data key="content">광 교 신도시 택지 공급 수지
분석주택 분양 </data>
</field>
</document>
</documentml>
```

<그림 4> 출력된 XML 파일의 분류 결과 부분

3. 평가

XML 결과 파일의 내용과 분류된 소기능, 주관부서의 주요업무와 BRM을 연계해 비교함으로써 분류 결과의 적합 여부를 판단하였다. 주관부서의 주요 업무는 경기도 홈페이지에 게시된 것을 활용하였다. 생성된 결과 파일 중, 100개의 샘플을 추출한 후 적합 여부를 Yes와 No로 구분해 표시하였다. 그 결과

경기도 웹사이트 게시글의 자동 분류 시스템은 83%의 적합성을 보였다. 하지만, 게시판의 특성상 특정 기간에 해당 기관 내에 쟁점 이슈가 발생하게 되면, 동일 글이 반복 된다는 맹점이 있기 때문에, 해당 글의 적합 여부가 전체 분류 결과에도 영향을 미칠 수 있다는 점을 고려해야 할 필요가 있다. 본 연구의 경우에도 '광교 신도시', '에콘힐', '도청사'와 같은 쟁점 이슈가 여러 건 포함되어 있었고, 해당 글의 적합 여부가 전체 시스템의 적합성 결과에 영향을 미치고 있는 것으로 나타났다.

따라서 좀 더 객관적인 성능을 파악하기 위해 샘플 데이터에 중복 글의 포함을 최소화할 수 있는 방안을 고려하여 재평가하였다. 게시판의 쟁점 이슈를 파악해 해당 이슈를 제외한 데이터만 출력하도록 수정하였고, 생성된 결과 파일은 681개였다. 그 중, 파일을 50개씩 랜덤으로 샘플링, 총 5번을 평가하였다. 그 결과 자동분류의 적합성은 52%로 감소하였다.

4. 결 론

본 연구는 공공게시판의 게시글을 공문서화하기 위해 정부기능연계모델(BRM)을 분류체계로 하여, 자동 분류 시스템을 구현하였다.

랜덤으로 추출된 게시글 100개의 분류 결과에 대한 적합성 평가 결과 83%의 적합성을 보이는 것으로 확인되었지만 쟁점 사안에 대한 중복 게시글을 제외한 250개의 게시글에 대한 평가 결과, 적합성이 52%로 감소하였다. 따라서 본 연구의 목적을 달성하기 위해서는 적합성을 좀 더 높여야 할 필요가 있다.

본 연구에서 제안한 텍스트 분류 시스템의 분류 성능에 영향을 미치는 요소에는 BRM 전처리 수준, 한글 형태소 분석기의 성능, BRM의 단계별 가중치 등이 있다. 본 연구에서는 형태소 분석의 한계를 피하기 위해 단어 matching에서 exact matching이 아닌 포함 여부(contain 함수 사용)를 적용하였으나 만족스러운 결과를 얻지 못하였다.

분류 과정에서 BRM의 각 단계별 항목 수에 따라 가중치를 부여하였는데, 엄밀하게는 각 단계별로 각 항목의 개수가 다르기 때문에, 각 단계뿐 아니라 항목별로도 가중치를 다르게 부여했어야 한다는 점에서 한계가 있다.

향후 연구에서는 적합성을 높이기 위해 맥락에 따라 달라지는 단어의 의미를 변별해내고, 동일한 의미의 유사어를 BRM의 단어와 동일 단어로 인식할 수 있는 시스템을 구현할 예정이다. 행정 용어 시소러스의 적용을 그 방안으로 고려하고 있다. 또한, 본 연구의 시스템을 보다 발전시키기 위해 SVM(Support Vector Machine) 분류기를 적용할 예정이다. SVM 분류기를 적용하면 문서의 내용을 바탕으로 미리 정의된 범주를 문서에 부여함으로써 문서를 자동 분류하는 자동 문서 범주화가 가능해지므로(정영미, 임해영 2000), BRM의 특정 단계에 대하여 자동 문서 범주화를 실시한 뒤 단위과제를 분류하면 좀 더 정확한 분류가 가능할 것으로 전망된다.

참고문헌

정영미, 임혜영. (2000). SVM 분류기를 이용한 문서 범주화 연구. 정보관리학회지, 17(4), 229-248.

정부혁신지방분권위원회. (2003). 참여정부의 전자정부 로드맵.

한성산. (2010). 경기도 기록관의 기록관리체계의 현황과 개선방안. 석사학위논문. 중부대학교. 기록관리전공.

Barry, R. (2004). Web sites as recordkeeping & recordmaking systems. Information Management Journal, 38(6), 26-30.

Sprehe, J. T. & McClure, C. R. (2005). Lifting the burden. Information Management Journal, 39(4), 47-52.