맥락정보를 이용한 기록 자동분류시스템 설계*

Design of Automatic Records Classification System Using Contextual Information

장 지 숙(Ji-Sook Jang)**
이 해 영(Hae-Young Rieh)***

목 차

〈초 록〉

기록학에서의 분류는 기록 자체의 내용보다는 기록이 생산되고 활용되는 맥락에 초점을 둔다. 본 연구에서는 업무활동이 반영된 기록을 업무활동 분석에 기반하여 구축된 분류체계에, 개별 기록의 내용이 아닌 기록의 집합적 맥락을 중심으로 자동분류 할 수 있는 기록 자동분류시스템을 설계하였다. 기 분류된 기록집합체뿐 아니라 분류체계와 시소러스를 분류기준으로 같이 구축하여 상호보완 할 수 있도록 설계하였으며, 분류대상기록의 범주를 할당한 후 바로, 분류된 기록의 맥락정보를 실시간으로 분류기준에 반영할 수 있는 방안도 포함하였다. 설계된 기록 자동분류시스템은 맥락정보의 품질에 따라 시스템의 성능이 좌우되는 한계가 있지만, 이를 통해 맥락정보를 제대로 충실하게 남길 수 있도록 유도하는 역할을 할 수 있다고 판단되었다.

주제어: 기록분류, 자동분류, 분류체계, 맥락정보, 업무활동 분석, 기록범주화

<ABSTRACT>

The classification in the Records and Archives Sciences focuses on the contextual information in producing and utilizing records rather than their contents. This study aimed at designing an automatic records classification system to enable an automatic classification focusing on the aggregation of the context of records rather than the contents of individual record in the classification scheme, structured on the basis of business activities analyses for records reflecting the business activities. The automatic records classification system was designed to have mutual supplements by constructing the classification scheme and thesaurus together as the classification reference, as well as the aggregation of records that have been already classified. Additionally included are plans to apply the classified contextual information of records to the classification reference on the real-time base right after the category assignment of records to be classified. Although there are limitations as the designed system depends on the quality of the contextual information, it is considered that the system could lead to ensure that the contextual information of records should be more substantial.

Keywords: records classification, automatic classification, classification scheme, contextual information, business activities analyses, records categorization

---

* 본 연구는 장지숙의 석사학위 논문(2009)인 「맥락정보를 이용한 기록 자동분류시스템 설계」를 요약・수정한 것임.
** 명지대학교 기록정보과학전문대학원 석사과정 졸업(withsoop@hanmail.net)(제1저자)
*** 명지대학교 기록정보과학전문대학원 교수(hyrieh@mju.ac.kr)(교신저자)
■ 접수일자 2009년 5월 16일 ■ 수정일자 2009년 6월 10일 ■ 게재확정일자 2009년 6월 18일

1. 서 론

업무의 전산화와 기록의 중요성에 대한 인식의 확산은 기록의 양적 증가를 촉진했으며, 많은 양의 기록을 빠르고 안정적으로 찾기 위해 기록의 체계적 관리가 매우 중요해졌다. 분류는 기록을 보다 더 잘 관리할 수 있는 기반구조를 제공한다.

전문적 지식과 기술을 갖춘 업무담당자나 기록관리자에 의한 분류는 그 품질이 보장될 수 있지만, 증가하는 기록물을 처리하는데 많은 시간과 노력이 필요하다. 또한 복잡한 정보를 올바르게 파악하여 분류의 품질을 유지하기 위해서는 더 많은 인력과, 이들에 대한 교육과 시간적 투자 등이 필요할 것이다. 이에 요구되는 시간 및 비용은 한계가 있기 마련이고, 이러한 한계를 극복하기 위해 컴퓨터를 통한 기록의 자동분류 도입을 검토해 볼 수 있다.

다양한 주제범주의 문서에 대해, 사람의 노력은 최소화하면서 마치 사람이 분류하듯이 분류 품질을 높이고자 하는 문서 자동분류에 대한 연구는 다양한 각도로 꾸준히 진행되어 왔다. 여러 연구 성과를 토대로 다양한 시스템들이 개발되어 왔으나, 아직 기록학 분야에서 기록의 자동분류에 대한 연구는 미비한 실정이다. 그러므로 기록 자동분류의 가능성을 찾는 작업은 시도해볼만한 유용한 과제이다.

기록의 생산과 동시에 등록 및 분류가 이루어질 수 없는 환경에서 생산된 기록은 시간과 인력 부족 등의 원인으로 시스템으로 획득되지 못해 관리되지 못하는 사례가 많다. 이러한 기록들을 시스템에 등록하거나 분류, 또는 재분류하고자 할 때 기록 자동분류기술을 이용한다면 매우 유용할 것이다. 즉, 그동안 없었던 분류체계를 새로 구축한 후 이전에 생산된 대량의 기록물을 분류하고자 할 때, BPR[1] 등이 도입된 후 사용하던 분류체계를 재구축할 때, 또는 혁신적 변화로 인해 기존 분류체계에 속해있던 기록물을 새 분류체계로 마이그레이션 할 필요가 있을 때 등이 이러한 경우에 해당될 것이다.

구체적 예로는 정부기관의 경우, 단위과제 중심의 정부기능분류체계(BRM, Business Reference Model)[2]와 단위업무 중심의 기록물분류기준표가 병존하는 상황을 해결하고자 할 때나 종이기록을 포함한 구기록을 시스템에 등록하고 분류체계에 배치하고자 할 때 기록의 자동분류기술은 유용할 것이다. 또한 이전에 기록관리 대상이 아니었던 이메일이나 웹사이트 등을 기록관리 대상에 포함시키고자 할 때에도 기록 자동분류기술은 많은 가능성을 제시할 수 있을 것이다.

업무와 분류체계, 기록의 존재이유 등 맥락에 대해 잘 알고 있는 업무담당자가 기록을 생산함과 동시에 분류체계에 배치할 수 있도록 시스템이 구축되어 있는 경우에도 기록의 자동분류기술을 도입하게 되면 업무의 편의성을 가져다 줄 수 있으며, 분류 시간을 단축할 수 있을 것이다. 업무담당자가 어느 범주에 기록을 편

1) Business Process Reengineering, 업무재설계: 경쟁 우위 확보를 위해 기업의 핵심부문에서 비용, 품질, 서비스, 스피드와 같은 요인의 획기적인 향상을 이룰 수 있도록 프로세스를 근간으로 업무 시스템을 근본적으로 재설계하여 극적인 성과를 추구하는 것을 말한다.
2) 본 논문에서는 이후 BRM으로 전개한다.

입시켜야 할 지 확신한다고 하더라도 계층적 분류체계에서 해당 범주를 찾는 데에는 시간과 노력이 필요하다. 특히 팝업창을 통해 분류체계를 찾아야 한다면 이는 더 번거로운 작업이 된다. 자동분류기술을 도입하여 할당될 범주를 자동으로 제시해 준다면 업무담당자는 확인만으로 분류가 가능할 것이다. 만약 분류 정확도가 떨어진다면 가장 적합한 순서로 범주를 복수 추천하여 그 중에서 선택하는 방법을 생각해 볼 수 있다. 이러한 과정은 1차로 자동분류기술을 이용해 분류한 것을 토대로 사람이 검증작업을 거쳐 최종 분류하는 것이므로 분류의 정확도를 높일 수도 있을 것이다.

본 연구를 시작할 때에는 지금까지 연구된 문서 자동분류에 대한 이론과 기법을 기록의 자동분류에 그대로 차용하는 것을 검토하였고, 알려진 문서 범주화 모델 중 어느 모델의 성능이 기록을 자동분류하는데 가장 우수한지 실험 및 분석을 통해 밝히고자 하였다. 그러나 지금까지 연구되어 온 문서 자동분류방법을 그대로 기록의 자동분류에 적용하기에는 여러 간극이 존재했다. 가장 근본적인 차이점은 일반적인 문서의 분류는 문서의 내용에 초점을 두지만 기록학에서의 분류는 기록 자체의 내용보다는 기록이 생산되고 활용되는 맥락에 초점을 둔다는 사실이었다. 또한 기록의 분류체계는 기관 고유의 특성을 지닌 업무기능과 활동에 기반한 기능분류체계를 주로 사용하기 때문에 표준 분류체계를 마련하기 어렵다. 따라서 특정 기관의 기록을 대상으로 기록 분류의 성능을 측정하여 가장 좋은 범주화 모델을 밝히는 것은 별로 의미가 없는 것으로 판단되었다.

이에 따라 본 연구의 목적은 업무활동이 반영된 기록을 업무활동 기반의 분류체계에 기록학적 관점, 즉 개별 기록의 내용이 아닌 기록의 집합적 맥락을 중심으로 자동분류하는 방법과 절차를 상세히 설계하는데 두었다. 본 연구에서는 분류 성능을 측정하지는 않았다. 분류규칙과 분류구조가 기관마다 다르기 때문에 기록의 분류 성능 또한 기관마다 다를 수밖에 없으며, 따라서 특정 기관의 기록을 대상으로 측정한 분류 성능이 우수하다하여 다른 기관에서도 우수한 성능을 보일 것이라고 일반화할 수 없다고 판단되었기 때문이다.

2. 기록의 자동분류를 위한 맥락정보

기록학에서 기록 분류의 초점은 기록 자체의 내용(content)보다는 기록이 생산되고 활용되는 맥락(context)이며, 기록은 '기록이 무엇에 관한 것인가(what it is about)'보다는 '기록이 왜 존재하는가(why it exists)'를 기초로 분류해야 한다(National Archives of Australia 2003a, 7). 이에 본 연구에서도 기록의 자동분류에 필요한 분류기준으로 이러한 맥락을 활용하고자 하였으며, 맥락의 관계를 분석하여 최종 분류하는 방법을 찾고자 하였다. 이를 위해 맥락정보 유형과 관계 등을 기능분류체계, 메타데이터, 기능시소러스로 나눠 분석한 후 이를 기록 자동분류시스템 설계에 반영하였으며, 이해를 돕기 위해 정부에서 현재 사용 중인 분류체계와 분류시스템, 업무관리시스템, 기록관리시스템에서 관리하고 있는 맥락정보를 샘플로 활용하였다.

분류체계는 기록을 분류하고 기술하는데 기

준이 된다. 분류체계를 통해 우리는 동일한 활동 혹은 연관 있는 활동에서 생성되는 기록간의 관계를 설명할 수 있으며 하나의 기록을 보다 큰 기록집합체의 어느 위치에 놓아야 하는지 결정할 수 있다(Elizabeth Shepherd & Geoffrey Yeo 2003, ch3). Schellenberg(1956)는 기록은 기능의 결과이며 기능과 관련하여 이용되기 때문에 기능에 따라서 분류되어야 한다고 주장하였으며, Bearman(1994)은 출처 의미의 중심은 조직구조가 아니라 기능에 있다고 제기하였다. 이에 따라 ISO 15489 및 호주, 영국, 미국 등 여러 나라 정부에서 업무 및 기능에 입각한 분류체계 개발을 제안하고 있으며, 우리나라에서도 BRM을 정립하여 운영하고 있다.

기록은 내용뿐 아니라 기록의 구조, 기록 형식, 그 기록을 구성하는 요소들 사이의 관계가 원래대로 남아 있어야 하며, 기록이 생산되고 획득되고 활용되는 업무 맥락이 기록에 명백히 드러나야 한다. 이것은 메타데이터를 통해 가능하다. 기록관리를 위한 구조화된 메타데이터는 업무활동과 기록집합체를 연결시켜 기록을 분류체계 속에 배치하는데 참조될 것이다. 행정부의 업무관리시스템의 과제관리카드와 문서관리카드는 메타데이터의 집합체라 할 수 있다. 과제관리카드는 단위과제 하위에 있는 기록철 구성단위로서 이를 통해 조직 및 개인이 수행할 과제(업무)를 BRM에 따라 체계적으로 분류하여 관리할 수 있으며 업무의 맥락정보를 파악할 수 있다. 문서관리카드는 과제와 관련된 의사결정을 위해 작성하는 개별 문건으로 보고서 작성, 보고경로 지정 및 처리과정에서 제기된 의견, 관련 자료 등을 기록하여 관리하는 카드로 행정의 투명성과 책임성을 증거 해준다.

기능시소러스는 기록과 다른 업무 정보를 관리할 수 있도록 도와주는 분류 도구(National Archives of Australia 2003b, 6)로 분류체계의 용어를 활용하여 기능시소러스로 확장할 수 있다. 기능시소러스는 분류체계와 함께 기록에 대한 종합적인 색인과 검색체계를 제공하며, 기록의 제목을 정하는 데에도 기준을 제시해 줄 수 있다(설문원 2003, 463). 기능시소러스 구축은 업무 기능과 활동 및 주제 사이의 논리적인 관계를 분석하여 상위기능어, 하위기능어, 관련기능어 관계를 구축하고, 비우선어(nonpreferred term)를 우선어(preferred term)로 연결시키는 것과 해당 기능어의 범위를 설명하는 것이 포함된다. 기능시소러스를 구축하는 업무의 대부분은 동의어 관계를 연결 짓는 일이다. 즉, 우선어와 비우선어를 연결시키는 것이다(National Archives of Australia 2003b, 14-16).

3. 기록 자동분류방법 개관

일반적인 문서의 자동분류에서는 주제에 따라 양질의 학습문서를 충분히 확보할 수 있지만, 기록 자동분류에 있어 가장 어려운 점은 실제 기록관리 환경에서 자동분류기술을 이용하고자 할 때에 학습기록을 초기에 충분히 확보하는 것이 어렵다는 것이다. 기관 고유의 특성을 지닌 업무기능과 활동에 기반한 기능분류체계를 사용하는 기록의 분류에서는 외부의 기록을 이용하여 학습할 수 없다. 기관자체내 실무에서 이미 분류된 기록을 학습에 이용하고자 할 때도 각 범주별로 기 분류된 기록의 수가 일정하지 않으며, 범주를 새로 만든 경우 기 분류

된 기록은 전혀 존재하지 않는다. 이것은 범주별로 기 분류된 기록에서 얻을 수 있는 분류기준 정보에 심한 불균형을 초래하여 자동분류의 품질에 영향을 주게 된다.

이에 따라 기 분류된 기록집합체뿐 아니라 각 분류범주의 개념 정보와 범주간 관계 정보를 잘 나타내주는 분류체계와 시소러스를 기 분류된 기록집합체와 함께 분류기준으로 구축하여 상호보완 역할을 할 수 있도록 기록 자동분류시스템을 설계하였다. 분류체계나 시소러스의 정보만을 분류기준으로 이용할 경우 각 범주를 구분해 줄 수 있는 정보량이 부족하며, 기능어의 변화나 새로운 용어 등을 많이 반영하지 못하게 되므로 기 분류된 기록집합체를 분류기준으로 같이 이용했고, 분류대상 기록을 분류한 직후 분류된 기록의 맥락정보를 실시간으로 분류기준에 반영하였다.

문서의 내용을 입력받아 분류하는 일반적인 문서 자동분류와는 달리 기록의 맥락정보를 이용해 기록을 자동분류해 주는 시스템을 구축하고자 한 본 설계에서는 기록의 맥락정보를 나타내주는 여러 요소 중 분류기준에 적합한 요소를 선택하는 일이 분류의 성능에 결정적인 역할을 한다. 따라서 분류기준을 구축하는 첫 번째 단계로 분류기준 요소 선택 단계를 두고 BRM과 기능시소러스, 맥락정보의 집합체라 할 수 있는 과제관리카드와 문서관리카드 각각에서 분류기준 요소로 적합한 맥락정보를 제시하였다. 정부의 업무관리시스템이나 기록관리시스템에서 관리하고 있는 맥락정보를 중심으로 제시하였지만 설계하는 기록의 자동분류 방법이 정부기록만을 대상으로 하지는 않는다.

분류에 이용하고자 하는 맥락정보는 업무담당자나 기록관리자가 규칙이나 업무지침에 따라 기록의 내용, 구조, 맥락을 가장 잘 표현하고자 전문성을 발휘한 정보이다. 이것을 기록의 자동분류에 이용하므로 다음과 같은 이점이 있다. 자동분류에서 분류범주 구분에 결정적인 단서가 되는 핵심 용어 즉 자질 선택 단계에서 간단한 방법으로 양질의 자질을 보장할 수 있다. 이미 업무담당자나 기록관리자 등 전문가에 의해 작성된 맥락정보를 기준정보로 이용하기 때문에 1차적 자질 선택은 이미 이루어진 것이다. 따라서 자동분류시스템에서 수행하는 자질선택 방법은 용어의 출현빈도만으로 2차 자질을 선택한다. 1, 2차에 걸쳐 선택된 자질은 전문가에 의해, 중요도에 의해 이중으로 선택된 양질의 자질로 분류의 중요한 기준이 될 수 있다.

분류기(classifier)의 범주화 모델에서는 학문적 연구보다는 현실상황을 반영한 실무적 이용을 고려하여 자동분류 알고리즘을 설계하였다. 본 연구에서 제안하는 기록 분류기의 범주화 모델은 분류대상기록이 각 분류범주의 맥락정보를 얼마나 가지고 있으며 그 맥락정보는 분류범주를 얼마나 대표하는가를 계산하여 할당될 범주를 결정하는 통계적 기법을 사용한다. 여기서 사용하는 통계적 기법은 TF(Term Frequency, 용어출현빈도)가 높을수록 해당 용어가 범주를 대표할 확률이 높다는 가설에 근거한 것이다. 하지만 TF는 각 범주별 맥락정보의 길이가 길거나, 기 분류된 기록의 수가 많을 경우 통계량 수치가 커지게 된다. 이런 편차를 줄이기 위해 각 범주별 최대 빈도수로 각 자질별 빈도수를 나누어 정규화 하였다. 여기서 또 하나 고려할 사항이 있다. 모든 범주에 고르게 출현하는 용어도 있기 때문에 TF만 가지고

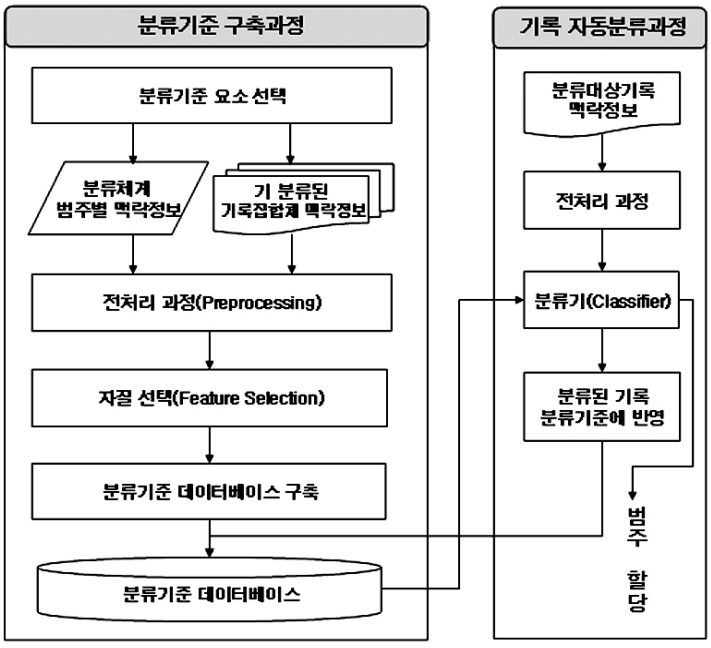
는 정확한 분류를 수행하기에 부족하다. 이 점을 보완하고자 범주간의 분리도가 높은 자질에 높은 가중치를 주는 ICF(Inverted Category Frequency, 역범주빈도)를 같이 이용하여 분류 알고리즘을 설계하였다.

TF와 ICF를 이용한 기록의 분류는 분류모델의 알고리즘이 간단하고 수행속도가 빠르지만 분류의 정확도에는 한계가 있을 수 있다. 이러한 한계는 여기에서 설계된 기록 자동분류시스템이 업무담당자나 기록관리자 등 전문가에 의해 작성된 맥락정보를 분류기준정보로 이용함으로써 해소될 수 있을 것이다.

끝으로, 본 기록 자동분류시스템은 정부의 업무관리시스템이나 기록관리시스템과 마찬가지로 분류체계의 최하위 범주 즉, 단위과제에만 기록을 분류하는 것을 전제로 설계하였다.

4. 기록 자동분류시스템 설계

기록 자동분류시스템은 〈그림 1〉과 같이 크게 분류기준 구축과정과 기록 자동분류과정으로 구성한다. 분류기준 구축과정에서는 기록의 맥락정보를 나타내주는 여러 요소 중 분류기준에 적합한 요소를 선택한 후, 선택된 분류체계의 범주별 맥락정보와 기 분류된 기록집합체의 맥락정보에서 각 범주를 대표할 수 있는 자질을 추출하여 벡터로 표현하는 작업을 수행한다. 이렇게 표현된 범주 벡터는 분류과정에서 분류대상 기록의 범주를 결정하는 기준이 된다. 기록 자동분류과정에서는 분류대상 기록의 맥락정보에서 추출한 용어리스트를, 분류기(classifier)를 이용해 범주별 분류기준과 매칭시켜 범주별 적합도를 구한 후, 적합도가 가장 높은 범주로 기록을 분



〈그림 1〉 기록 자동분류시스템 구성도

류한다. 이렇게 분류된 기록의 맥락정보는 분류기준에 반영되어 범주별 분류기준을 강화한다.

4.1 분류기준 구축과정

분류기준 구축과정은 맥락정보를 이용해 분류에 필요한 각 범주별 분류기준을 만드는 과정으로, 크게 네 단계로 진행된다. 첫 번째 단계는 분류기준으로 사용하기에 적합한 요소를 분류체계 범주별 맥락정보와 기 분류된 기록집합체 맥락정보에서 선택하는 단계이다. 두 번째 단계에서는 이렇게 선택된 맥락정보의 내용에서 형태소 분석기를 이용해 명사만을 추출한 후 불용어는 제거하고, 추출된 내용어와 출현빈도를 리스트로 만든다. 다음 단계에서는 이 중 각 분류범주의 특징을 잘 반영할 수 있는 자질(feature)만을 선택한다. 자질 선택 기법으로 TF(용어출현빈도)를 이용한다. 마지막 단계에서는 이렇게 선택된 자질리스트를 기록 분류과정에서 분류기준으로 이용할 수 있도록 데이터베이스로 구축하게 된다.

4.1.1 단계 1: 분류기준 요소 선택 단계

분류기준을 구축하는 첫 번째 단계로 해야 하는 일은 기록의 맥락정보를 나타내주는 여러 요소 중 분류기준으로 사용하기에 적합한 요소를 선택하는 일이다. 분류기준 요소 선택 단계는 문서의 내용을 입력받아 분류하는 일반적인 문서 자동분류에서는 필요치 않은 단계다. 그러나 기록의 맥락정보를 분류기준으로 이용하여 기록을 자동분류해 주는 시스템을 구축하고자 하는 본 설계에서는 반드시 필요한 단계다.

본 설계에서는 예시로 우리나라 기록관리의 대표성을 지닌 정부에서 현재 관리하고 있는 맥락정보를 분류기준으로 활용하였다. 분류기준으로 고려해 볼 수 있는 맥락정보는 국가기록원에서 제시한 「기록관리 메타데이터 표준-현용·준현용 기록물 용」 필수요소를 참고하여, BRM과 기능시소러스, 맥락정보의 집합체라 할 수 있는 과제관리카드와 문서관리카드에서 찾아보았다.

먼저 BRM에서는, 모든 계층에서 공통적으로 관리되고 있는 분류체계(기능)ID, 분류체계(기능)명, 상위기능명, 기능설명을 분류기준으로 선택하였다. 선택된 분류체계ID는 각 범주별 분류기준의 기본키(Primary Key)로 설정한다. 분류체계의 최하위 범주 즉, 단위과제에만 기록을 분류하기로 전제하였기에 여기서 분류체계명은 단위과제명이 되며 하위기능명은 분류기준 요소로 선택하지 않고 상위기능명만을 선택한다. 상위기능명은 기능시소러스에서도 선택할 수 있지만 기능시소러스가 구축되지 않은 경우가 많기 때문에 BRM에서 선택하는 것으로 한다. BRM에서 관리하고 있는 단위과제에 대한 다양한 정보는 과제관리카드에서 좀 더 상세한 정보를 제공하므로 과제관리카드에서 제공하는 맥락정보를 이용한다.

분류기준은 기능시소러스에서도 찾을 수 있다. 기능시소러스 내에 우선어인 분류체계명에 설정되어 있는 관계어들은 상위어, 하위어, 관련어, 비우선어, 범위주기가 있다. 이중 상위어와 하위어는 기능분류체계의 계층구조를 이용할 수 있으므로 기능시소러스에서는 관련어와 비우선어를 분류기준 요소로 이용한다. 특히 비우선어는 우선어인 분류체계명과 등가관계이며 우선어로 안내하는 역할을 수행, 우선어

의 범위를 명확히 하는데 도움이 되는 중요한 정보이다. 범위주기는 적용 범위를 명확하게 구분할 수 있도록 해 주기 때문에 중요한 정보이다. 하지만 사람이 아닌, 기계를 이용하는 자동분류에서는 포함하는 기능과 포함하지 않는 기능을 용어만을 이용해 명확하게 구별하기 어렵다. 포함하는 기능과 포함하지 않는 기능을 따로 입력받아 관리하는 것을 고려해볼 수 있지만 이것 또한 가장 작은 단위의 명사로 형태소 분석을 실시할 경우 제외시켜야 하는 기능어를 명확히 걸러내기가 어렵다. 따라서 아쉽게도 범위주기는 분류기준 요소로 포함시키지 않았다.

기능분류체계인 BRM과 기능시소러스를 따로 언급했지만 기능분류시스템에는 두 가지가 함께 포함되어야 한다. 각 기능별로 이미 구축되어 있는 기본정보, 속성정보, 유관정보, 업무편람 외에 관련정보에 시소러스를 추가하여 상위어, 하위어, 관련어, 비우선어, 범위주기를 관리해야 한다. 기능분류체계에 기능시소러스가 추가된다면 업무와 기록의 연계와 기록의 분류 및 활용에 있어 정확성과 용이성을 증가시킬 수 있다.

과제관리카드에서 관리하는 맥락정보에서는 표제부 중 관리정보(공개여부, 열람범위)를 제외한 내용 및 취지, 과제이력, 과제담당자, 내부관계자가 분류기준 요소로 적합하다. 관리정보는 기록을 분류체계에 편입시키고 난 이후에, 편입된 분류체계에 부여된 관리정보를 그대로 사용하는 것으로 기록 등록 이후의 속성이므로 제외한다.

기록은 단독으로 존재하지 않고 항상 논리적 집합체로 존재한다. 동일한 활동에서 산출된 기록은 서로 밀접하게 연관되어 있다. 따라서 특정 업무의 계층적 구조로 이미 분류된 기록집합체의 맥락정보를 분류기준으로 활용한다면 아직 분류되지 않은 기록을 적절한 기록집합체에 연결하여 분류체계로 배치하는데 많은 역할을 할 수 있다. 이미 분류된 기록집합체의 맥락정보를 분류기준으로 제공하기 위해 마지막으로 문서관리카드에서 제목, 검색어, 정보출처, 문서취지를 분류기준 요소로 선택한다.

분류기준 요소 중 BRM 분류체계에서 분류체계명, 시소러스에서 비우선어, 문서관리카드에서 제목과 검색어는 가중치를 부여한다. 왜냐하면 이 요소의 맥락정보들은 이미 전문가에 의해 맥락을 가장 잘 드러내 줄 수 있는 결정적인 핵심용어만이 선택되어 표현된 정보이기 때문이다.

지금까지 분류기준 요소로 선택한 맥락정보를 정리하면 〈표 1〉과 같다.

위에서 분류기준 요소로 선택한 맥락정보만이 정답은 아니다. 이 요소들은 현재, 정부에 구축된 시스템과 기록관리 메타데이터 표준 중 필수요소 등을 참고하여 기록을 자동분류할 때 기준정보로 적합하다고 판단되는 요소들을 선정한 것이다. 기관에 따라 기능시소러스가 구축되지 않은 곳도 있을 것이고, 기록관리 메타데이터 표준 중 필수요소뿐만 아니라 선택요소까지 시스템으로 잘 구축된 곳도 있을 것이다. 또한 정부의 업무관리시스템이나 기록관리시스템을 중심으로 설명하지만 설계하는 기록의 자동분류 방법이 정부기록만을 대상으로 하지는 않는다.

정리하면 분류기준 요소 선택 단계는 해당 기관의 맥락정보를 대상으로 기록관리자가 분류기준 정보로 적합하다고 판단되는 요소를 선택하는 단계이다. 기관 실정에 따라 선택할 수 있지만, 선택한 맥락정보가 범주를 구별해 주

〈표 1〉 맥락정보 중 분류기준으로 선택한 요소

출 처		분류기준 요소	비 고
분류체계 범주별 맥락정보	분류체계	분류체계(기능)ID	각 범주의 기본키
		분류체계(기능)명	*
		상위기능명	최상위 기능명~현 레벨 전 기능명
		기능설명	
	시소러스	상위어	시소러스가 미구축된 경우가 많기 때문에 분류체계에서 요소 선택
		하위어	
		관련어	
		비우선어	*
	과제관리카드	내용 및 취지	
		과제이력	
		과제담당자	부서, 성명
		내부 관계자	부서, 성명
기 분류된 기록집합체	문서관리카드	제목	*
		검색어	*
		정보출처	
		문서취지	
		기록 생산자	부서, 성명

(*: 맥락정보에 가중치 부여)

는 정보가 아니라 잡음을 발생시키는 원인이 되는 정보가 아닌지 주의를 기울여야 한다. 기록 자동분류시스템 구축 시에는 기관에서 관리하는 맥락정보 리스트를 보여주고 그 중에서 분류기준 요소와 가중치 부여 여부를 선택할 수 있도록 인터페이스를 제공한다.

맥락정보 중 분류기준 요소와 가중치 부여 여부를 결정할 때 적용한 원칙은 〈표 2〉와 같다.

4.1.2 단계 2: 전처리 단계

전처리(preprocessing) 단계에서는 단계 1에서 선택한 분류기준 요소의 내용에서 형태소 분석기를 이용하여 명사만을 추출한다. 추출된 용어 중에는 별다른 정보를 주지 못하는 지시대명사와 같은 불용어(stop word)들이 포함되어 있으므로 이들 불용어는 제거하고 추출된 내용어와 출현빈도를 리스트로 만든다.

〈표 3〉은 정부의 업무관리시스템인 온-나라 시스템의 올바른 활용을 위한 지침서(행정자치부 2007)에서 '분류체계명'과 '내용 및 취지' 작성 사례를 인용하여 형태소 분석을 실시한 결과이다. 형태소 분석 시 가중치를 부여하겠다고 선택한 맥락정보와 그렇지 않은 것을 각각 분석한 후, 가중치를 부여한 분류기준 요소에서 추출된 내용어는 출현빈도를 두 배(출현빈도×2)로 계산한다. 출현빈도는 설계한 기록 자동분류시스템에서 핵심적인 역할을 하기 때문에 출현빈도를 높이는 가중치 부여 방법은 의미가 있다. 〈표 3〉에서 보면 '혁신'의 출현빈도는 '1'이지만 '×2'를 해 출현빈도가 '2'가 됨을 알 수 있다. 최종 결과는 두 그룹으로 나누어 분석한 결과를 합쳐 분류체계ID를 기준으로 추

〈표 2〉 분류기준 요소 선택 시 적용 원칙

원 칙	설 명
대표성	해당 맥락정보는 분류범주나 기록을 대표할 수 있다.
중요성	해당 맥락정보는 필수적으로 관리되어야 한다.
핵심성	해당 맥락정보 내용은 핵심용어만을 사용하여 표현된다.
획득가능성	해당 맥락정보는 현장에서 공통적으로 관리하는 맥락정보로, 획득될 가능성이 높으며 기록을 분류체계에 배치하기 전에 획득 가능하다.
명확성	해당 맥락정보 내용은 사람이 아닌 기계를 이용해 분석할 경우에도 혼돈을 일으키지 않을 정도로 명확하다
충분성	해당 맥락정보 내용은 분류범주나 기록을 충분히 설명할 수 있다.
비교성	해당 맥락정보 내용을 비교하면 분류범주를 유추할 수 있다.

〈표 3〉 형태소 분석 결과 예

원 문	형태소 분석 결과			
	추출 내용어	빈도	추출 내용어	빈도
[분류체계(과제)명] 10개 혁신도시 건설 추진	10	2(1×2)	건설	2(1×2)
	혁신	2(1×2)	추진	2(1×2)
	도시	2(1×2)		
[내용 및 취지] • 목적 - 공공기관 지방이전을 계기로 수도권과 지방의 균형발전을 도모하고, 지역 발전 촉진 • 내용 - 혁신도시개발예정지구 지정 - 혁신도시 개발계획 및 세부 실시계획 수립 - 혁신도시 건설공사 착수 - 부동산 투기방지 및 난개발 방지대책 수립·시행 • 기대효과 - 수도권 인구집중 완화를 통해 수도권 질적발전의 계기 마련 - 지방의 자립역량 강화	도시	3	목적	1
	발전	3	부동산	1
	수도권	3	세부	1
	지방	3	시행	1
	혁신	3	실시	1
	개발	2	역량	1
	계획	2	예정	1
	방지	2	완화	1
	수립	2	이전	1
	강화	1	인구	1
	건설	1	자립	1
	계기	1	지구	1
	공공	1	지역	1
	공사	1	지정	1
	균형	1	질	1
	기관	1	집중	1
	기대	1	착수	1
	난개발	1	촉진	1
	내용	1	통해	1
	대책	1	투기	1
	도모	1	효과	1
	마련	1		

출 내용어와 총 출현빈도를 리스트로 만든다.

전처리 과정에서의 Input과 Output, 알고리즘은 〈표 4〉와 같다. 전처리 알고리즘을 수행하면 '분류체계ID, 내용어, TF(용어출현빈도)'로 구성된 리스트가 결과로 생성된다. 그중 한 범주 예를 〈표 5〉에 표시했다. 앞의 예를 분류체계ID가 '050302'라 가정하고 진행시킨 결과 중 일부를 표시하였다.

4.1.3 단계 3: 자질 선택 단계

자질(feature) 선택 단계는 특정 분류범주 구분에 결정적인 단서가 되는 핵심 용어를 선택하는 단계이다. 일반적인 정보검색시스템에서는 문서에 나타난 대부분의 단어를 사용하지만 자동분류에서는 기계학습을 적용할 수 있도록 차원을 축소하는 것이 일반적이다. 단계 2를 거쳐 도출된 내용어를 모두 자질로 선택할 경

〈표 4〉 전처리 알고리즘

```
Input: 단계 1에서 선택한 분류기준 요소
Output: 전처리_리스트(분류체계ID, 내용어, TF)

Begin
    1. 분류기준으로 선택된 요소의 값(맥락정보)을 분류체계ID를 기준으로 가져오기
    2. 범주별 형태소 분석
    Do While Not EOF3)              '결과적으로 범주 수만큼 반복4)
       2.1 가중치가 부여된 맥락정보
           원칙에 따라 형태소 분석
           분석결과 TF × 2
       2.2 나머지 맥락정보
           원칙에 따라 형태소 분석
       2.3 "2.1.수행결과"와 "2.2.수행결과" 합쳐 리스트 생성
           전처리_리스트(분류체계ID, 내용어, TF)
    Loop
End
```

〈표 5〉 전처리 후 결과 리스트 예

분류체계ID	내용어	TF	분류체계ID	내용어	TF	분류체계ID	내용어	TF
050302	도시	22	050302	균형	10	050302	수도권	3
050302	혁신	20	050302	발전	8	050302	수립	2
050302	지방	16	050302	자립	6	050302	예정	1
050302	개발	13	050302	역량	6	050302	지구	1
050302	건설	10	050302	10	4	050302	통해	1

(분류체계ID: '050302', 분류체계명: '10개 혁신도시 건설 추진' 가정)

3) End Of File.

4) 알고리즘에서 주석은 주석내용 앞에 " ' "를 붙이고 달았다. 따라서 " ' " 이후 표현은 알고리즘 내용이 아니고 그에 대한 설명이다.

우 분류범주 기준을 구축하는 시간은 물론이고 분류 과정에서도 너무 많은 시간을 소모하게 된다. 따라서 분류범주의 특징을 잘 반영할 수 있는 자질만을 분류기준으로 구성하고 나머지는 제외시켜야 한다. 자질 선택은 전체적인 성능에 영향을 미치는 중요한 과정이다.

본 논문에서 제안하는 자질 선택 기법은 간단하지만 분류의 중요한 단서가 될 수 있다. 이미 업무담당자나 기록관리자 등 전문가에 의해 작성된 맥락정보를 분류의 기준정보로 이용하기 때문에 1차적 자질 선택은 이미 이루어진 것이라 볼 수 있다. 따라서 이 단계에서는 용어의 출현빈도만으로 2차 자질을 선택한다. 1, 2차에 걸쳐 선택된 자질은 전문가에 의해, 중요도에 의해 2중으로 선택된 양질의 자질로 분류의 중요한 단서가 될 수 있다.

단계2에서 얻은 용어의 출현빈도를 기준으로 일정 빈도수 미만인 것은 제외한다. 출현빈도가 낮은 용어는 특정 분류 범주를 대표할 만한 충분한 정보가 되지 못하기 때문에 자질로 선택할 경우 분류 성능에 기여하지 못하고 오히려 방해가 되는 잡음으로 작용한다.

제거기준으로 삼을 빈도수 선택은 맥락정보의 충실도 및 기 분류된 기록의 수에 따라 관리자가 선택할 수 있도록 한다. 디폴트로 TF(용어출현빈도)가 2 미만인 것으로 하고 만약 기관의 맥락정보의 내용이 충실하게 기술되어 있거나, 기 분류된 기록의 수가 많을 경우 TF가 3 미만인 용어를 제거하는 것도 고려해 볼 수 있다. 기록 자동분류시스템을 기관에서 실제 업무에 사용할 때, 디폴트인 'TF〈2 제거' 또는 'TF〈3 제거' 등을 선택한 후 각각 나오는 분류결과의 정확도를 모니터한 후 선택하는 것도 하나의 방법이다.

자질 선택 과정에서의 Input과 Output, 알고리즘은 〈표 6〉과 같다. 자질제거기준TF의 디폴트를 2라고 설정하고, 단계 2 수행결과인 전처리_리스트를 입력받아 내용어TF가 2미만인 내용어를 제거한 결과가 〈표 7〉이다. 자질로 선택된 결과이므로 내용어를 자질로 바뀌어 표시했다.

4.1.4 단계 4: 분류기준 데이터베이스 구축 단계

분류기준 데이터베이스 구축 단계는 단계 3까지 진행된 결과로 생성된 자질 리스트를, 실

〈표 6〉 자질 선택 알고리즘

```
Input: 단계 2 수행결과인 전처리_리스트(분류체계ID, 내용어, TF), 자질제거기준TF=2
Output: 자질_리스트(분류체계ID, 자질, TF)

Begin
    For i = 1 To Length(전처리_리스트) '전처리_리스트 길이(개수)만큼 반복
        If 내용어TF < 자질제거기준TF Then
            리스트에서 해당 내용어(분류체계ID, 내용어, TF) 제거
        End If
    Next i
End
```

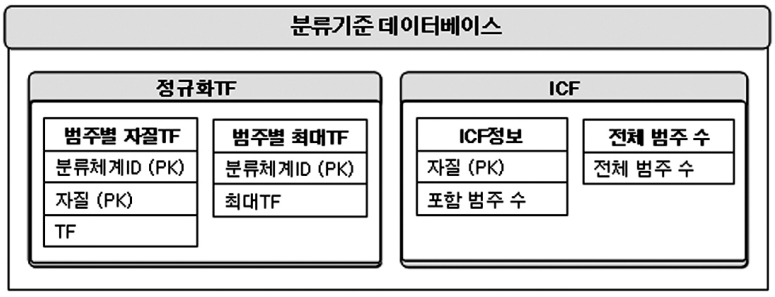
〈표 7〉 자질 선택 후 결과 리스트 예

분류체계ID	자질	TF	분류체계ID	자질	TF	분류체계ID	자질	TF
050302	도시	22	050302	균형	10	050302	수도권	3
050302	혁신	20	050302	발전	8	050302	수립	2
050302	지방	16	050302	자립	6			
050302	개발	13	050302	역량	6			
050302	건설	10	050302	10	4			

제 기록을 분류할 때 분류기준으로 이용할 수 있도록 데이터베이스로 구축하는 단계이다. 단순히 자질리스트를 데이터베이스에 저장하는 것뿐만 아니라 분류기(classifier)에 설계된 알고리즘을 실행하는데 필요한 정보를 미리 준비해 둔다.

분류기에서 사용할 알고리즘은 기록 자동분류 과정에서 상세히 설명하겠지만, 분류기에서 필요로 하는 정보를 미리 준비해 두기 위해 간단히 언급하자면, 분류기에서는 정규화TF(용어출현빈도)와 ICF(역범주빈도)를 이용해 기록을 분류한다. 정규화TF는 각 자질별 빈도수를 각 범주별 최대 빈도수로 나누어 계산하기 때문에 자질의 TF와 함께 범주별 최대TF 정보도 필요하다. 그리고 ICF를 계산하기 위해서는 각 자질을 포함하고 있는 범주 수와 전체 범주 수가 필요하다. 분류기에서 기록의 분류 속도를 빠르게 하기 위해서 정규화TF와 ICF 계산에 필요한 위의 정보를 미리 계산하여 〈그림 2〉와 같이 범주별 자질리스트와 함께 분류기준 데이터베이스로 구축한다.

분류기준 데이터베이스 구축 과정에서의 Input과 Output, 알고리즘은 〈표 8〉과 같다. 단계 3 수행결과인 자질리스트를 입력받아 데이터베이스를 구축한 결과 예를 〈표 9〉에 표시했다. 그동안 분류체계ID가 '050302', 분류체계명이 '10개 혁신도시 건설 추진'인 하나의 범주에 대해서만 예를 들어왔는데, 분류단계 때 비교설명을 위해 분류체계ID가 '010503', 분류체계명이 '경력 개발 프로그램 지원'인 범주를 하나 더 추가한 후, 그동안 진행했던 단계를 거쳐 분류기준 데이터베이스로 구축한 예를 〈표 9〉에 포함했다.



〈그림 2〉 분류기준 데이터베이스 스키마

〈표 8〉 분류기준 데이터베이스 구축 알고리즘

Input: 단계 3 수행결과인 자질_리스트(분류체계ID, 자질, TF)
Output: 분류기준 데이터베이스

Begin
1. 자질_리스트(분류체계ID, 자질, TF) DB에 저장
2. 범주별 최대TF 찾아 DB에 저장
3. 자질별 포함 범주 수 계산한 후 DB에 저장
4. 전체 범주 수 DB에 저장
End

〈표 9〉 분류기준 데이터베이스 구축 예

범주별 자질TF

분류체계ID	자질	TF	분류체계ID	자질	TF
050302	도시	22	050302	수립	2
050302	혁신	20	010503	개발	25
050302	지방	16	010503	전문	22
050302	개발	13	010503	경력	18
050302	건설	10	010503	지원	15
050302	균형	10	010503	프로그램	15
050302	발전	8	010503	역량	10
050302	자립	6	010503	양성	6
050302	역량	6	010503	향상	6
050302	10	4	010503	인력	5
050302	수도권	3	010503	수립	2

범주별 최대TF

분류체계ID	최대TF	분류체계ID	최대TF
050302	22	010503	25

전체 범주 수

전체 범주 수
20

ICF정보

자질	포함 범주 수	자질	포함 범주 수	자질	포함 범주 수
10	1	수도권	2	지방	2
개발	4	수립	6	지원	6
건설	3	양성	3	프로그램	4
경력	2	역량	4	향상	4
균형	2	인력	2	혁신	3
도시	4	자립	3		
발전	4	전문	3		

(분류체계ID: '050302', 분류체계명: '10개 혁신도시 건설 추진', 분류체계ID: '010503', 분류체계명: '경력 개발 프로그램 지원'을 예로 구축)

4.2 기록 자동분류과정

기록 자동분류과정은 분류대상기록을 가장 적합한 범주로 자동할당하는 과정이다. 분류기준 구축과정이 분류체계ID를 기준으로 작업이 진행되었다면 기록 자동분류과정은 이 분류체계ID를 찾는 과정이다. 기록 자동분류과정은 세 단계를 거치게 되는데, 첫 번째 단계는 분류대상기록의 전처리 단계로, 분류대상기록의 맥락정보 내용을 분류기준 구축과정에서와 마찬가지로 형태소 분석기를 이용해 내용어만을 추출한 후 출현빈도와 함께 리스트를 만든다. 두 번째 단계는 분류기를 이용한 기록자동분류단계로, 생성된 분류대상기록의 전처리 리스트에 있는 내용어와 일치하는 자질을 분류기준 데이터베이스에서 찾은 다음, 일치하는 각 자질이 해당 범주에 기여하는 정도를 정규화TF와 ICF(역범주빈도)를 이용해 계산한 후 내용어 출현빈도를 곱하여 범주별 분류적합도를 계산한다. 그리고 범주별 분류적합도 계산이 끝나면 분류적합도 비교과정을 거쳐 가장 큰 분류적합도를 가진 범주로 기록을 분류한다. 마지막 단계에서는 이렇게 분류된 기록의 맥락정보를 실시간으로 분류기준에 반영하여 분류기준을 강화한다.

4.2.1 단계 1: 분류대상기록의 전처리 단계

'분류기준 구축과정'에서 했던 전처리 방식과 원칙을 그대로 적용한다. 단지 전처리 대상만 다르다. '분류기준 구축과정'에서 행했던 전처리는 분류체계와 시소러스, 과제관리카드에서 분류기준 요소로 선택된 분류체계 범주별 맥락정보와 기 분류된 기록의 문서관리카드에서 분류기준 요소로 선택된 맥락정보를 대상으로 했다면, '기록 자동분류과정'의 전처리는 분류대상기록의 맥락정보를 입력받아 진행한다. 분류대상기록의 맥락정보 모두를 대상으로 하는 것이 아니라 기 분류된 기록의 맥락정보 중 분류기준 요소로 선택된 즉, 문서관리카드 중 제목, 검색어, 정보출처, 문서취지, 기록 생산자(부서, 성명)에 해당하는 내용만 형태소 분석기에 입력된다.

분류대상기록의 전처리 과정에서의 Input과 Output, 알고리즘은 〈표 10〉과 같다. 분류대상기록의 맥락정보 내용을 가져와 전처리 알고리즘을 수행하면 '내용어, TF(용어출현빈도)'로 구성된 리스트가 결과로 생성된다. 생성된 리스트 예는 〈표 11〉과 같다. 생성된 결과리스트가 분류대상기록을 대표하여 다음단계에서 범주별 분류기준과의 비교분석을 통해 특정 범주로 할당된다.

4.2.2 단계 2: 분류기를 이용한 기록 자동분류 단계

기록 자동분류 단계는 분류기(classifier)를 이용해 분류대상기록을 가장 적합한 분류체계의 범주로 할당하는 단계이다. 분류기의 범주화 모델을 어떻게 설계하느냐에 따라 이 단계의 알고리즘은 달라진다. 본 연구에서는 분류대상기록을 어느 범주로 할당되는 것이 적합한지 판단하는 기준으로, 분류대상기록이 각 분류범주의 맥락정보를 얼마나 가지고 있는가 하는 점을 고려하여 통계적 기법을 사용하였다.

통계적 기법은 TF(용어출현빈도)가 높을수록 해당 용어가 범주를 대표할 확률이 높다는 가설에 근거한 것이다. TF가 높다는 것은 분류

〈표 10〉 분류대상기록의 전처리 알고리즘

```
Input: 기록의 맥락정보 중 선택된 분류기준 요소
Output: 분류대상기록전처리_리스트(내용어, TF)

Begin
    1. 분류기준으로 선택된 분류대상기록의 맥락정보 가져오기
    2. 가중치가 부여된 맥락정보
       원칙에 따라 형태소 분석
       분석결과 TF × 2
    3. 나머지 맥락정보
       원칙에 따라 형태소 분석
    4. "2.수행결과"와 "3.수행결과" 합쳐 리스트 생성
       분류대상기록전처리_리스트(내용어, TF)
End
```

〈표 11〉 분류대상기록의 전처리 후 결과 리스트 예

내용어	TF	내용어	TF	내용어	TF
개발	4	균형	2	인력	3
경력	6	수립	1	자립	1
교환	1	역량	4	효과	1

기준 요소로 선택된 맥락정보에서 자주 언급되었다는 것이고, 이러한 용어들은 그만큼 해당 범주에서 중요하게 취급되는 용어라 할 수 있다. 하지만 TF는 각 범주별 맥락정보 내용의 길이와 기 분류된 기록의 수에 영향을 받는다. 각 범주별 맥락정보의 길이가 길거나, 기 분류된 기록의 수가 많을 경우 통계량 수치가 커진다. 이런 편차를 줄이기 위해 각 범주별 최대 빈도수로 각 자질별 빈도수를 나누어 정규화하고자 한다.

또 하나 고려할 사항이 있다. 모든 범주에 고르게 출현하는 용어도 있기 때문에 TF만 가지고 특정 범주로 분류하기에는 허점이 있다. 예를 들면, 분류기준 요소로 선택한 분류체계의 상위기능명의 경우 형제끼리는 모두 같기 때문에 분류 시 분별력을 떨어뜨린다. 이러한 점을 보완하고자 각 범주간의 분리도를 나타내 줄 수 있는 ICF(역범주빈도)를 같이 이용하고자 한다. ICF는 범주간의 분리도가 높은 자질에 높은 가중치를 준다. IDF(역문서빈도)와 기본 원리는 같지만 IDF는 문서간의 분리도가 높은 자질에 높은 가중치를 주는 것이고, ICF는 범주간의 분리도가 높은 자질에 높은 가중치를 주는 것이라는 차이점이 있다. 즉 소수의 범주에 많이 출현한 자질에 대해서는 높은 가중치를 주고, 여러 범주에 고르게 출현한 자질에 대해서는 낮은 가중치를 주는 것이다.

단계 1의 결과인 분류대상기록전처리 리스트에 있는 내용어와 일치하는 각 범주별 자질을 분류기준 데이터베이스에서 찾아낸다. 그리

고 일치하는 각 자질이 해당 범주에 기여하는 정도를 다음과 같이 정규화TF와 ICF를 이용해 계산한다. i번째 범주에서 자질 k의 가중치 w_{ik}는 다음과 같다.

$$w_{ik} = \text{정규화}TF \times ICF$$
$$w_{ik} = \frac{f_{ik}}{Max_i f_{ik}} \times \log(\frac{C}{c_k})$$

여기서, $Max_i f_{ik}$: i번째 범주에 출현한 자질의 최대 빈도수

f_{ik}: i번째 범주에서 자질 k의 용어출현빈도 수

C: 전체 범주의 수

c_k: 자질 k를 포함하고 있는 범주의 수

사실 각 범주별 특정 자질의 가중치는 일반적인 문서의 자동분류에서는 학습과정에서 계산한다. 본 논문에서 제안하는 시스템의 단계로 보면 분류기준 구축과정에서 이미 계산이 되어야 했던 것이다. 그렇지만 본 시스템에서는 기록을 분류하기 직전에 계산하도록 하였다. 그 이유는 기록관리 실무에서는 분류(학습)기록을 초기에 충분히 확보하여 안정적인 분류기준을 만드는 것이 어렵기 때문에 분류기를 통해 분류된 기록을 실시간으로 분류기준에 반영하여 범주별 분류기준을 강화하기 위해서이다.

분류기준 구축과정에서 자질별 가중치를 모두 계산해 두면 분류단계에서 계산이 간단해지는 것이 사실이다. 그러나 이것은 분류기준이 확정적이고 불변일 경우에는 매우 유용하지만, 분류기준이 수시로 변경되는 경우에는 변경되는 기준과 관계있는 모든 것들에 대해 정규화TF와 ICF, 가중치를 일일이 다시 계산해야 하는 관계로 상당히 복잡하다. 특히 ICF의 경우 해당 범주뿐 아니라 많은 다른 범주에도 영향을 주기 때문에 관련된 범주를 모두 찾아 가중치를 다시 계산해야 한다.

이러한 문제를 해결하고자 분류직전에 분류대상기록전처리 리스트에 있는 내용어와 일치하는 각 범주별 자질에 대해서만 가중치를 계산하는 것으로 설계하였다. 이 경우 계산시간도 오래 소요되지 않으며 분류된 기록을 분류기준에 반영할 때도 간단하다. 분류된 기록의 맥락정보를 분류기준에 반영하는 방법은 다음 단계를 참고한다.

분류대상기록의 내용어와 일치하는 각 범주별 자질의 가중치를 계산한 다음, 이 가중치 값에 분류대상기록의 해당 내용어TF를 곱한다. 각 범주별로 분류대상기록의 내용어와 일치하는 자질의 수만큼 위의 계산을 반복한 후, 이 값을 모두 합하면 범주별 분류적합도가 계산된다. 그 다음 범주별 분류적합도를 비교하여 가장 큰 값을 가진 범주로 기록을 할당하면 된다.

내용어 리스트 $(c_1, c_2, c_3, \ldots, c_n)$로 구성된 분류대상기록 R이 i번째 범주로 분류될 적합도 $Fit(R \rightarrow i)$는 다음과 같이 계산한다.

$$Fit(R \rightarrow i) = \sum_{c=1}^{n} (w_{ik_c} \times tf_{ik_c})$$

여기서, w_{ik_c}: i번째 범주에서 기록 R의 내용어 c와 일치하는 자질 k의 가중치

tf_{ik_c}: 기록 R에서 자질 k와 일치하는 내용어 c의 출현빈도수

지금까지의 내용을 알고리즘으로 정리하면 〈표 12〉와 같다.

분류대상기록전처리__리스트 예 〈표 11〉과 〈표 9〉의 분류기준 데이터베이스 구축 예를 이용해 위 알고리즘을 차례로 수행해보겠다. 우선 [알고리즘 단계1]에 따라 분류기준 데이터베이스로부터 분류대상기록의 내용어와 일치하는 자질과 자질TF, 포함 범주 수, 범주 최대TF를 검색해온 결과와 전체 범주 수는 〈표 13〉과 같다.

〈표 12〉 분류기를 이용한 기록 자동분류 알고리즘

```
Input: 단계 1 수행결과인 분류대상기록전처리_리스트(내용어, TF)
Output: 분류대상기록 범주 할당

Begin
    1. 분류기준DB에서 분류대상기록 전처리_리스트의 내용어와 일치하는 범주별 자질TF(분류체계ID, 자질, TK)와
       관련된 ICF정보(자질, 포함 범주 수), 범주별 최대TF, 전체 범주 수 가져오기 '[알고리즘 단계1]
    2. 가져온 정보를 이용해 범주별 분류적합도 계산 '[알고리즘 단계2]
       For i = 1 To 범주 수 '상이한 분류체계ID 개수만큼 반복
           For k = 1 To(내용어 리스트와 일치하는 i번째 범주의 자질 수)
               i범주_k자질_분류적합도 = (i범주_k자질_TF / i범주_최대TF)
                                        × log(전체 범주 수 / k자질_포함 범주 수)
                                        × k자질_분류대상기록_내용어TF
               i범주_분류적합도 = i범주_분류적합도 + i범주_k자질_분류적합도
           Next k '다음 자질
       Next i '다음 범주
    3. 분류적합도가 가장 큰 범주로 할당 '[알고리즘 단계3]
End
```

〈표 13〉 기록 자동분류 [알고리즘 단계1] 실행결과 예

분류체계ID	자질	자질TF	포함 범주 수	범주 최대TF
50302	개발	13	4	22
50302	균형	10	2	22
50302	수립	2	6	22
50302	역량	6	4	22
50302	자립	6	3	22
10503	개발	25	4	25
10503	경력	18	2	25
10503	수립	2	6	25
10503	역량	10	4	25
10503	인력	5	2	25

전체 범주 수
20

[알고리즘 단계2]는 검색해 온 정보를 이용해 분류적합도를 계산하는 과정이다. 정규화TF와 ICF를 계산하여 두 곱으로 가중치를 구하고, 여기에 분류대상기록 내용어의 TF를 곱하면 특정범주의 특정자질에 대한 분류적합도가 계산되며, 이것을 모두 합하면 특정범주의 분류적합도가 된다. 〈표 11〉과 〈표 13〉의 결과를 이용해 분류체계ID가 '050302'인 '개발'을 예로 들어 분류적합도를 계산하면 다음과 같다.[5)]

정규화TF = 해당 범주 해당 자질TF / 범주 최대TF
= 13 / 22 = 0.59
ICF = log(전체 범주 수 / 해당 자질 포함 범주 수)
= log(20 / 4) = 1.61
가중치 = 정규화TF × ICF
= 0.59 × 1.61 = 0.95
분류적합도 = 가중치 × 분류대상기록 내용어TF
= 0.95 × 4 = 3.80

위 계산을 [알고리즘 단계1] 실행결과에 모두 적용하면 〈표 14〉와 같다.

마지막 과정인 [알고리즘 단계3]은 [알고리즘 단계2]의 결과를 바탕으로 분류적합도가 가장 큰 범주로 분류대상기록을 할당한다. 위 결과로 보면 분류체계ID가 '050302'인 범주의 분류적합도는 '8.22', 분류체계ID가 '010503'인 범주의 분류적합도는 '20.2'가 된다. 따라서 다음과 같이(내용어, TF)로 표현할 수 있는 '분류대상기록(R) = {(개발, 4), (경력, 6), (교환, 1), (균형, 2), (수립, 1), (역량, 4), (인력, 3), (자립, 1), (효과, 1)}'은 분류체계ID가 '010503', 분류체계명이 '경력 개발 프로그램 지원'인 범주로 할당된다.

〈표 14〉 기록 자동분류 [알고리즘 단계2] 실행결과 예

(소수점 둘째자리 미만 반올림)

분류체계ID	자질	자질TF	정규화TF	ICF	가중치	내용어TF	분류적합도
050302	개발	13	0.59	1.61	0.95	4	3.80
050302	균형	10	0.45	2.30	1.04	2	2.08
050302	자립	6	0.27	1.90	0.51	1	0.51
050302	역량	6	0.27	1.61	0.43	4	1.72
050302	수립	2	0.09	1.20	0.11	1	0.11
050302	범주 분류적합도						8.22
010503	개발	25	1.00	1.61	1.61	4	6.44
010503	경력	18	0.72	2.30	1.66	6	9.96
010503	역량	10	0.36	1.61	0.58	4	2.32
010503	인력	5	0.20	2.30	0.46	3	1.38
010503	수립	2	0.08	1.20	0.10	1	0.10
010503	범주 분류적합도						20.2

5) 모든 계산은 소수점 둘째자리 미만 반올림 적용.

4.2.3 단계 3: 분류된 기록의 맥락정보, 분류기준에 반영 단계

이 단계는 단계 2에서 분류된 기록의 맥락정보를 분류기준 데이터베이스에 반영하는 단계이다. 본 시스템에서는 분류된 기록을 실시간으로 분류기준에 반영하여 처음 분류기준 구축 시 범주별로 기 분류된 기록의 수에 의해 존재했던 분류기준의 불균형을 실무적으로 해소하고자 한다. 이를 위해 분류대상기록의 분류시점에 기록의 맥락정보를, 할당된 범주의 분류기준을 강화하는데 이용한다. 그러면 분류기록이 늘어날 때마다 분류기준에 대한 신뢰성이 증가하고 분류의 정확도가 향상될 것이다. 또 다른 이점은 용어의 변화나 새로운 용어의 출현을 분류기준에 곧바로 반영하여 분류에 기여하도록 한다는 점이다.

단계 1의 수행결과인 분류대상기록의 전처리리스트(내용어, TF)를 단계 2의 수행결과로 할당된 범주의 기준정보로 반영하기 위해서는 다음과 같은 과정을 거친다. 우선 분류대상기록의 내용어가 할당된 분류범주의 자질로 이미 있는지 확인하여, 만약 있다면 자질TF에 내용어TF를 합산한다. 자질TF가 변경되었기 때문에 분류범주의 최대TF가 바뀔 가능성이 존재한다. 따라서 기존의 최대TF와 비교하여 더 크다면 할당된 분류범주의 최대TF를 갱신한다. 만약 분류대상기록의 내용어가 할당된 분류범주의 자질로 존재하지 않는다면 분류대상기록의 내용어를 할당된 분류범주의 자질로 추가하는 작업을 진행한다. 이때 자질로 선택될 자격이 있는지 자질제거기준TF를 이용해 확인해야 한다. 만약 내용어TF가 자질제거TF 미만일 경우에는 자질로 선택될 수 없다. 자질로 선택된 내용어는 분류기준 데이터베이스에 분류범주의 자질로 추가한다. 이때도 분류범주의 최대TF가 바뀔 가능성을 확인해야 한다. 특정 범주에 자질이 추가되었기 때문에 ICF정보도 갱신해야 한다. 추가한 자질이 ICF정보에 있다면 포함 범주 수만 하나 증가시키면 되지만, 만약 없다면 포함 범주 수를 1로 하여 자질과 함께 ICF정보에 새로 추가한다. 이러한 과정을 Input과 Output을 포함해 알고리즘으로 정리한 것이 〈표 15〉이다. 상세한 알고리즘을 제공했기에 갱신된 분류기준 데이터베이스 예는 생략한다.

〈표 15〉 분류기록의 맥락정보 분류기준에 반영 알고리즘

```
Input: 단계 1 수행결과인 분류대상기록전처리_리스트(내용어, TF)
       단계 2 수행결과인 분류대상기록에 할당된 분류범주, 자질제거기준TF
Output: 갱신된 분류기준 데이터베이스

Begin
    For i = 1 To Length(분류대상기록전처리_리스트) '내용어 개수만큼 반복
        If 할당된 분류범주에 분류기록의 내용어가 이미 자질로 존재 Then
            자질TF = 자질TF + 내용어TF
            If 자질TF 〉 범주별 최대TF Then
                범주별 최대TF = 자질TF '범주별 최대TF 갱신
            End If
        Else '할당된 분류범주에 분류기록 내용어가 자질로 미존재
```

```
                If 내용어TF 〉= 자질제거기준TF Then
                    범주별 자질TF(분류체계ID, 자질, TF)에(분류체계ID, 내용어, TF) 추가
                    If 자질TF 〉 범주별 최대TF Then
                        범주별 최대TF = 자질TF '범주별 최대TF 갱신
                    End If
                    If ICF정보에 추가한 자질이 존재 Then
                        포함 범주 수 = 포함 범주 수 + 1
                    Else 'ICF 정보에 추가한 자질 미존재
                        ICF정보(자질, 포함 범주 수)에(추가한 자질, 1) 추가
                    End If
                End If
            End If
        Next i '다음 내용어
    End
```

5. 결 론

본 연구에서는 업무활동이 반영된 기록을 업무활동 분석에 기반하여 구축된 분류체계에, 개별 기록의 내용이 아닌 기록의 집합적 맥락을 중심으로 자동분류 할 수 있는 기록 자동분류시스템을 설계하였다.

본 연구 과정에서 설계된 기록 자동분류시스템은 맥락정보의 품질에 따라 시스템의 성능이 좌우된다. 이 점은 한계로 지적될 수도 있겠지만, 반대로 맥락정보를 제대로 충실하게 남길 수 있도록 유도하여 서로 상승작용을 이끌 수도 있을 것이다. 기록관리에서는 기록 자체를 온전하게 보존하는 것 못지않게 맥락정보를 제대로 남기는 것 또한 중요하다. 따라서 본 시스템의 분류성능을 모니터하여 정확도가 낮은 범주에 대해 범주간의 분리도에 문제가 없는지 분류체계를 점검하고, 맥락정보가 명확하지 않거나 내용이 빈약하지는 않은지 점검하는 계기로 삼아 개선을 유도할 수 있을 것이다.

기록 자동분류에 대한 본 연구는 이제 겨우 걸음마를 뗀 정도다. 이 논문에서 제시하고 있는 설계를 바탕으로 기록 자동분류시스템을 실제로 구축하여, 각기 다른 기관에서 실제 기록을 대상으로 테스트와 보완을 거쳐 신뢰할 수 있는 기록 자동분류시스템을 완성해 실무에 도움이 될 수 있도록 하는 것이 앞으로 주어진 과제이다.

요즘 추세는 기록관리시스템, 컨텐츠관리시스템, 지식관리시스템 등을 통합하여 운영하는 것이다. 업무활동을 기반으로 하는 기능분류체계를 축으로, 본 연구에서 설계한 맥락정보를 이용한 자동분류기술에 각 시스템에서 생산되는 맥락정보를 대체한다면 각각에서 생산되는 정보와 기록을 상호 연계해서 이용하는 것이 가능할 것이다. 그렇게 된다면 업무수행 중 획득한 다양한 지식정보와 기록을 체계적으로 축적하고 공유하여 활용할 수 있는 포탈시스템으로의 통합이 가능해지므로 이로 인한 시너지 효과는 상당할 것으로 판단된다. 따라서 본 연구를

확대하는 것 또한 향후 과제로 남길 수 있겠다.

본 연구를 통해 아직까지 기록학 분야에서 도입하지 않고 있었던 기록의 자동분류에 대한 가능성을 찾아보았다. 적용 가능한 영역에서 이 기록 자동분류시스템이 실제로 구축되고 활용될 수 있다면 업무담당자나 기록관리자의 업무 효율성과 편의성 향상에 기여할 수 있을 것이다.

참 고 문 헌

고영중, 서정연. 2002. 문서관리를 위한 자동문서범주화에 대한 이론 및 기법. 「정보관리연구」, 33(2): 19-32.

국가기록원. 2007. 「기록관리 메타데이터 표준-현용·준현용 기록물 용」.

방선이. 2006. 「개념 기반 시소러스를 이용한 귀납적 문서 범주화의 성능 향상」, 전북대학교 컴퓨터통계정보학과 박사학위논문.

설문원. 2003. 조선총독부 기록물을 위한 기능분류 체계 개발. 「정보관리학회지」, 20(1): 457-488.

이영숙. 2001. 「계층적 분류체계를 위한 자동분류 기법에 관한 연구」, 연세대학교 문헌정보학과 석사학위논문.

행정자치부. 2006. 「단위과제 관련정보 작성방법(정부기능분류시스템)」.

행정자치부. 2006. 「온라인 정부업무관리시스템 사용자 따라하기」.

행정자치부. 2007. 「온-나라 시스템의 올바른 활용」.

행정자치부. 2007. 「정부기능분류시스템 운영지침」.

Bearman, David. 1994. *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh: Archives & Museum Informatics.

Shepherd, Elizabeth & Geoffrey Yeo. 2003. *Managing Records: a Handbook of Principles and Practice*London: Facet Publishing.

KS X ISO/TS 23081-2:2008 문헌정보-기록관리과정-기록메타데이터-제2부: 개념과 실행쟁점 표준(안), 산업자원부 기술표준원.

ISO 23081-1:2006 Information and documentation-Records management processes-Metadata for records-Part 1:Principles.

Lubbes, R. Kirk. 2001. "Automatic Categorization: How It Works, Related Issues, and Impacts on Records Management." *Information Management Journal*, 35(4) (Oct 1, 2001): 38-42.

Lubbes, R. Kirk. 2003. "So you want to implement automatic categorization?" *Information Management Journal*, 37(2) (Mar/Apr 2003): 60-68.

National Archives of Australia. 2003a. "Overview of Classification Tools for Records

Management," [cited 2008.10.01]. 〈http://www.naa.gov.au/Images/classifcation%20tools_tcm2-1030.pdf〉.

National Archives of Australia. 2003b. "Developing a Functions Thesaurus," Guidelines for Commonwealth Agencies, [cited 2008.10.01]. 〈http://www.naa.gov.au/Images/developing-a-thesaurus_tcm2-916.pdf〉.

Schellenberg, Theodore R. 1956. *Modern Archives: Principles and Techniques* ; 이원영 역. 2002.「현대기록학개론」. 서울: 진리탐구.

Yang, Yiming & Jan O. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, 412-420.

Yang, Yiming. 1999. "An Evaluation of Statistical Approaches to Text Categorization," *Journal of Information Retrieval*, 1: 69-90.