논문 2011-6-9

연관규칙 마이닝을 활용한 뉴스기사 키워드의 연관성 탐사

Discovering News Keyword Associations Using Association Rule Mining

김한준*, 장재영**

Han-joon Kim, Jaeyoung Chang

요 약 현재 대부분의 웹포털 사이트는 인기도 또는 중요도가 높은 키워드를 제공하는 서비스가 제공되고 있는데, 구체적으로 태그 클라우드 형태와 연관 검색 서비스와 같은 사용자 친화형 서비스를 지원하고 있다. 하지만 일반적으로 뉴스기사는 날짜와 분야별로 기사들이 분류되어 있기에, 사용자는 카테고리별로 나누어진 기사를 읽을 수만 있을 뿐 그 기사와 연관된 다른 기사를 쉽게 찾아보지는 못한 실정이다. 또한 연관 검색어 서비스도 사용자가 검색한 입력 내용을 기반으로 연관성 정도를 분석하기에 충분한 객관성을 보장하지 못하고 있다. 본 논문에서는 기존의 태그 클라우드 방식에서 좀 더 나아가 축적된 뉴스 기사로 부터 검색 키워드와 밀접히 연관된 키워드를 추출하여 제공하는 기사 검색 방식을 제안한다. 제안 기법은 기본적으로 연관규칙 마이닝을 이용하여 키워드 연관성을 추출하게 되며, 뉴스기사 특성을 반영하여 문장 내부에 존재하는 키워드에 한정하여 연관성을 추출한다. 연관된 키워드 집합을 이용하여 키워드와 가장 밀접한 기사를 검색할 뿐만 아니라, 연관 키워드간의 관계성을 보여줌으로써 뉴스 기사들 속에 숨겨진 연관정보의 탐색을 가능하게 한다.

Abstract The current Web portal sites provide significant keywords with high popularity or importance; specifically, user-friendly services such as tag clouds and associated word search are provided. However, in general, since news articles are classified only with their date and categories, it is not easy for users to find other articles related to some articles while reading news articles classified with categories. And the conventional associated keyword service has not satisfied users sufficiently because it depends only upon user queries. This paper proposes a way of searching news articles by utilizing the keywords tightly associated with users' queries. Basically, the proposed method discovers a set of keyword association patterns by using the association rule mining technique that extracts association patterns for keywords by focusing upon sentences containing some keywords. The method enables users to navigate the space of associated keywords hidden in large news articles.

Key Words : Association Rule Mining, Keyword Analysis, Data Mining, Information Retrieval

Ⅰ. 서 론

기존의 웹포털 사이트가 제공하는 블로그(blog)나 검색(search) 기능을 포함한 인터넷 환경을 웹 1.0으로 본다면 개방적인 웹 환경을 기반으로 네티즌들의 정보공유와 참여가 가능하게 된 요즘의 인터넷 환경을 웹 2.0이라고 부른다. 웹 2.0 시대가 도래하면서 사용자들의 편의성을 추구하는 다양한 서비스 기술들이 등장하고 있다. 대표적으로 사이트에 새롭게 올라 온 글을 사용자가 원하는 정보만 볼 수 있도록 서비스화 되어있는 RSS(Really Simple Syndication)[1]나 사용자가 마음대로 지정해 놓은 단어가 중요한 정보로 검색 될 수 있는 태그(tag) 등을 예로 들 수 있다. 이러한 기술의 발달로 인해 최근 포털

*정회원, 서울시립대학교 전자전기컴퓨터공학부
**정회원, 한성대학교 컴퓨터공학과
접수일자 2011.10.11, 수정완료 2011.11.20
게재확정일자 2011.12.16

사이트의 뉴스 기사페이지는 인기가 있거나 중요도가 높은 내용들을 사용자가 질문하기 전에 미리 보여주고 정보의 동향을 제공해주는 태그 클라우드(tag cloud) 기능을 제공하고 있다[2][3]. 하지만 아직도 대부분의 뉴스기사 페이지들은 날짜와 분야별로 기사들이 나열되어 있으며 사용자는 카테고리별로 나누어진 기사를 읽을 수만 있을 뿐 그 기사와 연관된 다른 기사의 정보에 대해서 한눈에 알아 볼 수 있는 방법은 미흡한 실정이다.

본 논문에서는 기존의 태그 클라우드 방식에서 좀 더 나아가 축적된 뉴스 기사로 부터 검색 키워드와 밀접히 연관된 키워드를 추출하여 제공하는 검색 기법을 제안한다. 제안 기법의 핵심 요소인 연관검색 방식은 사용자가 기사 검색을 하였을 때, 키워드와 가장 밀접한 기사를 검색해 줄 뿐만 아니라, 키워드와 연관된 키워드들과 연관 정도를 보여준다. 이는 사용자가 검색한 기사의 내용을 파악할 뿐만 아니라 연관된 키워드에 대한 정보를 얻게 된다. 나아가 연관 기사들끼리의 관계성을 확인함으로써 자기가 원하는 정보이외에 관심이 없었던 기사의 연관까지 확인하여 정보 습득에 대한 시야를 더욱 넓혀 줄 수 있다.

기존의 일부 포털 사이트에서도 본 연구와 유사한 연관 검색 기능을 제공하고는 있다. 하지만 포털 사이트에서 제공되는 연관검색 기능은 사용자들이 입력한 키워드 간에 유사성이나 동일한 사용자에 의해 연속적으로 입력된 키워드들을 분석함으로써 키워드들 간의 연관성을 부여한다. 반면에 본 연구내용은 텍스트 형식의 뉴스 기사들을 분석하여 기사 내에 키워드들의 연관성을 분석하여 제공함으로써 객관성을 보장할 수 있다.

본 논문에서 제안하는 연관검색은 포털사이트에 축적된 뉴스 정보를 웹 크롤링(web crawling) 방식으로 수집한 후, 다양한 검색 기술들을 활용하여 키워드간의 연관성을 분석하였다. 본 연구에서는 키워드의 연관 정도를 분석하기 위해 통계 기반한 키워드 관련도(keyword association)를 정의하였다. 기존 방법과 다른 점은, 키워드 관련도를 하나의 문서 혹은 문서 집합에 대해서 고려했다는 것이며, 이는 기존 정보검색의 문서표현 기법인 TF-IDF 가중치 방식[4][5]과 연관규칙 탐색 방식인 Apriori 알고리즘[6]을 각각 개선한 형태를 취한다.

본 논문의 구성은 다음과 같다. 2장에서는 키워드 연관분석과 관련된 기존 연구들을 소개하고, 3장에서는 키워드 연관검색 기법을 소개한다. 4장에서는 제안 기법의 성능분석을 기술하고, 마지막으로 5장에서는 결론을 맺는다.

II. 배경연구

1. 연관규칙 마이닝

앞서 기술한 바와 같이, 본 연구에서는 문서집합내에 존재하는 단어들 간의 연관 패턴을 추출하기 위해 연관규칙 마이닝(association rule mining) 기술을 이용한다. 연관규칙 마이닝의 기본 개념은 데이터들에 대한 발생빈도를 기반으로 각 데이터 간의 연관관계를 도출하는 기법이다. 예를 들어, 연관규칙 '{A, B}→C' 의 의미는 상품 A, B를 구입하는 사람은 많은 경우 상품 C를 구입한다는 의미이다. 본 연구는 연관규칙 마이닝의 대표적인 알고리즘인 Apriori 알고리즘[6]을 토대로 한다.

일반적으로 연관규칙 마이닝은 2단계로 구성되는데, 1단계는 빈발항목집합(large itemset)을 구하는 것이고, 2단계는 1단계에서 생성된 빈발집합을 이용하여 연관규칙을 생성하는 것이다. '빈발항목집합'이란 해당 집합 내에 존재하는 항목들이 동시에 존재하는 확률인 지지도 값이 최소치(minimum support) 이상의 항목집합들로 구성된 것을 말한다. 그런데, 본 연구에서는 상품 집합 대신 키워드 집합이 마이닝 대상이기 때문에, 2단계 연관규칙 생성 단계는 고려할 필요가 없다. 즉, 1단계에서 구해진 빈발집합에 포함된 단어들이 모두 서로 관련이 있는 것으로 간주한다. 예를 들어, 빈발항목집합 {A, B, C}가 구해졌다면, 단어 A, B, C가 서로 연관된 것으로 처리된다.

그림 1은 기존의 Apriori 연관규칙 마이닝 알고리즘의 1단계 부분인 빈발집합(large itemsets)을 구하는 부분을 문서 데이터의 특성에 맞게 수정한 것이다. 우선 k개의 항목으로 구성된 후보항목집합(candidate itemset)을 생성하여, 이를 주어진 문서 데이터베이스와 비교하여 최소 지지도를 만족하는 빈발항목집합을 찾게 된다. 그림 1에서 보는 바와 같이, 빈발항목집합이 공집합이 될 때까지 반복하는 과정을 거치는데, k+1개의 항목을 가지는 후보 집합 C_{k+1}은 k개의 항목집합을 가지는 빈발항목집합 L_k를 셀프 조인(self-join) 연산을 수행함으로써 얻어진다. 이때 반복과정의 시간을 줄이기 위해서 Apriori 성질, 즉 빈발하지 않은 k개의 항목집합을 포함하는 항목집합 또한 빈발하지 않는다는 성질을 활용한다. 결과적으로, 생성된 빈발항목집합 L_k을 모두 결합한 결과가 연관

C_k : a candidate itemset with k items
L_k : a frequent itemset with k items
n : the number of transactions (i.e., documents)
begin
build L_1
for(k= 1; L_k != $\varnothing$: k++) do
    C_{k+1}= candidates generated from L_k
    for each c $\in$ C_{k+1}
        for each document d in database do
            if (c is contained in d) c.count++
        L_{k+1}= { c | c $\in$ C_{k+1} and c.count/n 〉= minimum support }
    end
return $\bigcup_k L_k$
end

그림 1. 기존 Apriori 연관규칙 마이닝 알고리즘
Fig. 1. Conventional Apriori association rule mining algorithm

된 키워드들의 집합이 되는 것이다.

2. 키워드 추출 및 연관관계 분석

키워드 추출은 정보검색(information retrieval), 문서분류(text categorization), 주제탐색(topic detection), 문서요약(document summarization) 등을 포함한 텍스트 마이닝(text mining) 분야의 연구에서 주요 속성(feature) 추출을 위해 사용되는 기술이다. 키워드간의 연관성을 분석하는 문제에 있어서도 키워드 추출이 선행되어야 하고, 키워드 추출과 관련된 기술들을 활용되기도 한다. 키워드의 가중치를 결정하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 기법을 활용할 수 있는데[7][8], 본 연구에서는 키워드 추출의 문제와 키워드 연관성 탐색 문제를 유사한 문제로 바라보았다.

문서 단어에 대한 연관규칙 마이닝에 대한 연구는 [9][10][11]에서 시작되었다고 할 수 있다. [12]에서는 통계 처리를 기반으로 하는 본 연구와는 달리 단어와 해당 텍스트간의 관련성을 분석하기 위해 자연어처리(natural language processing) 기술을 도입하였다. 이는 정확도를 높일 수 있는 가능성은 있지만, 자연어 처리 부하에 비해 그다지 성능이 높지 못하다. [11]은 문서내의 주요 단어와 주어 어구를 추출함으로써 단어와 텍스트를 연관시키는 기법을 제안하였다. [12]에서는 주어진 문서로부터 문서요약을 실현하기 위한 빈발 에피소드를 생성하는 범용 프레임워크를 제안하였으며, [13]에서는 텍스트 데이터에 대한 연관규칙을 시각화하는 시스템을 실현하였다.

Ⅲ. 키워드 연관성 분석

본 장에서는 두 키워드간의 관련성 정도를 정량화하기 위한 방안을 기술한다. 자연어처리는 구문 분석 및 의미 분석 과정을 통해 고유명사간의 관련성을 측정할 수 있으나[12], 정확도와 시간복잡도 문제를 감안하여 본 논문에서는 통계기반 방안을 제안한다.

우선 제안 기법의 아이디어를 도출하기 위해 정보검색에서 흔히 사용하는 TF-IDF 가중치 모델[4][5]을 분석해보자. 기본적으로 TF-IDF 가중치는 한 문서 내에서 한 단어의 가중치를 결정할 수 있는 모델로서, 이는 TF(term frequency)와 IDF(inverse document frequency)값을 곱한 것이다. 여기서 TF값은 한 문서 내에서의 특정 단어가 출현하는 횟수 (또는 이를 정규화한 값)를 의미하며 이는 한 문서 내에서의 그 단어의 중요도를 반영한다. 그리고 IDF값은 특정 단어가 출현하는 문서의 수(DF, document frequency)의 역수(또는 이를 정규화한 값)를 의미하며, 이는 전체문서집합에서 단어의 중요도를 반영한다. DF값이 큰 단어는 많은 문서에서 공통적으로 사용되는 단어이기 때문에 TF값이 크더라도 그 중요도를 낮춰져야 한다. IDF값은 TF값의 오류를 보

정한다는 측면에서 중요한 작용을 한다.

본 논문에서는 두 키워드간의 관련도를 측정하기 위해서 TF-IDF 가중치 모델과 유사한 접근방식을 취한다. 첫째, TF 측면에서 문서 내에서 두 키워드간의 관련도(Association Frequency, AF)를 정의한다. AF는 TF와 유사하게 한 문서 내부에 존재하는 두 개체간의 횟수를 정규화한 값이다. 이는 기본적으로 두 개체간의 관련성은 한 문장 내에서 두 개체가 공존하는 횟수로부터 도출될 수 있다.

다음은 문서 d_x내의 키워드 e_i와 e_j간의 관련도 $AF_{d_x}(e_i,e_j)$를 정의한 것이다.

$$AF_{d_x}(e_i,e_j)=\sum_{s_p\in d_x} SentenceAssoc_{d_x,s_p}(e_i,e_j) \quad (1)$$

$SentenceAssoc_{d_x,s_p}(e_i,e_j)$는 문서 d_x의 각 문장 s_p에 존재하는 키워드 e_i와 e_j의 관련도를 측정한 값으로서, 이는 한 문장에 존재하는 키워드들 간의 모든 조합수의 역수로 정의한다. 즉, 한 문장에서 다른 키워드들이 포함된 경우에 e_i와 e_j간의 관련도가 약화됨을 반영한 것이다. 예를 들어, 그림 2의 첫 문장에 존재하는 키워드의 수는 3개이므로, $SentenceAssoc_{d_x,s_p}(e_i,e_j)$ 값은 $1/{}_3C_2$ = 1/3이 된다. 이러한 SentenceAssoc값을 모두 더하여 한 문서 내에서의 Association Frequency (AF) 값으로 측정한다. 단, 한 문장에서 e_i 또는 e_j가 단독으로 출현하는 경우에는 합산하지 않는다.

둘째, IDF 측면에서, 전체 문서 집합 속에서 두 키워드간의 관련성을 정의하자. TF-IDF 가중치 모델에서 IDF 인자는 전체문서집합에 퍼져 있는 단어에 대한 가중치를 줄이기 위한 용도로 사용한다. 이에 반해서, 두 키워드간의 관계에 대한 정보가 전체 문서집합에 퍼져 있는 것은 그 관련성의 정도가 크다고 판단하는 것이 합리적이다. 따라서 본 연구에서는 DF 개념을 그대로 반영하여 키워드 간의 관련도를 계산한다.

결론적으로 위 두 가지 사항을 고려하여 모든 문서 내에서 두 키워드 e_i와 e_j의 관련도 $Assoc(e_i,e_j)$는 다음과 같다.

$$Assoc(e_i,e_j)=AF(e_i,e_j)*\{1+\log DF(e_i,e_j)\} \quad (2)$$

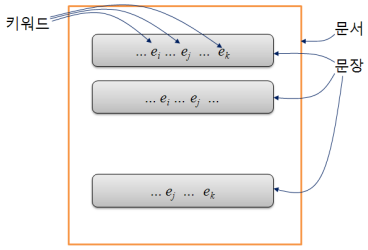


그림 2. 문서, 문장, 키워드의 개념도
Fig. 2. Conceptual diagram of sentences, sentences, and keywords

여기서, $AF(e_i,e_j)$는 각 문서에 대해 e_i와 e_j의 관련도를 모두 더한 값이며, $DF(e_i,e_j)$는 e_i와 e_j가 공존하는 문장이 존재하는 문서의 개수를 의미한다. 이 때 $DF(e_i,e_j)$의 값이 커지는 경우에 AF값의 신뢰성을 떨어뜨릴 수 있으므로, DF값에 log를 취하여 DF값의 증감에 따른 영향을 줄인다. DF를 고려한 이유는 e_i와 e_j가 하나의 문서에 집중적으로 나타남으로써 AF값을 증가시키는 것보다는 여러 문서에 걸쳐 동시에 출현하는 것이 관련성이 더 높은 것으로 간주하기 위해 설정한 요소이다.

추가적으로 e_i와 e_j의 관련도 $Assoc(e_i,e_j)$는 다음의 요소들을 고려할 수 있다. 우선 뉴스 기사는 대부분 기사 제목을 포함하고 있다. 따라서 제목에 키워드가 나타나는 경우는 키워드의 중요도가 당연히 높아짐으로 이에 대한 가중치를 부여할 수 있다. 또한 뉴스 기사 내에서 문서의 앞부분에 나타나는 키워드일수록 해당 기사와 관련성이 크다고 볼 수 있으므로 키워드들이 나타나는 문장이 문서의 어느 위치에 존재하는 가에 대해 가중치를 부과할 수도 있다. 이러한 추가 조치 방안은 기존의 대부분의 검색 시스템에 적용하고 있지만, 본 논문에서는 이와 관련한 구체적인 방안에 대해서는 고려하지 않았다.

그림 3은 기존 Apriori 연관규칙 마이닝 알고리즘을 토대로 식 (2)의 키워드 관련도 함수를 반영하여 개선한 것이다. 본 논문에서는 이를 Keyword_Apriori 알고리즘라고 명한다. 이 그림에서 보는 바와 같이, 빈발항목집합 L_k로부터 셀프조인 연산을 수행하여, 후보항목집합 C_{K+1}을 생성하고, C_{K+1} 집합에서 한 항목씩에 대하여 문서 데이터베이스와 비교하여, $AF_d(c)$ 함수값을 계산한다. 여기서 $AF_d(c)$ 함수는 식 (1)에서 보는 바와 같이, 두 개의 개체인자가 들어가는데 $AF_d(c)$는 후보항목집합 c에

C_k : a candidate itemset with k items
L_k : a frequent itemset with k items
begin
build L_1
for(k= 1; L_k != ∅ : k++) do
  C_{k+1}= candidates which are generated from L_k
  for each c ∈ C_{k+1}
    for each document d in database do //

$$AF_d(c) = \sum_{s \in d} SentenceAssoc_{d,s}(c)$$

$$AF(c) = AF(c) + AF_d(c)$$

    end

$$Assoc(c) = AF(c) * \{1 + \log DF(c)\}$$

  end
  L_{k+1}= { c | c ∈ C_{k+1} and Assoc(c) >= minimum support }
end
return $\bigcup_k L_k$

그림 3. Keyword_Apriori 연관규칙 마이닝 알고리즘
Fig. 3. Keyword_Apriori association rule mining algorithm

존재하는 항목들의 2쌍 조합들의 값을 계산함을 의미한다. 데이터베이스에 존재하는 모든 문서에 대하여 $AF_d(c)$ 값을 계산하여, 이를 합산한 결과가 $AF_d(c)$에 할당된다. 최종적으로 후보항목집합 내에 존재하는 각 항목 c에 대한 $AF(c)$ 값과 $DF(c)$ 값을 이용하여 $Assoc(c)$ 함수 값을 도출한다. 빈발항목집합 L_{k+1}은 $Assoc(c)$ 값이 미리 주어진 최소 지지도 (minimum support) 이상인 항목 c를 가려냄으로써 얻어진다.

Ⅳ. 실험 및 결과

1. 실험 방법 및 환경

본 논문에서 제안한 Keyword_Apriori 알고리즘이 기존 알고리즘보다 연관검색에 얼마나 효율적인지를 검증하기 위해 실험을 실시하였다. 실험은 2011년 3월부터 6월까지 연합뉴스의 정치관련 기사 1000여건을 대상으로 하였다. 이 뉴스기사들로부터 가장 많이 언급된 정치인 16인을 선별하고 이들과 가장 연관성이 있는 인물들을 기존 Apriori 알고리즘과 Keyword_Apriori 알고리즘을 이용해 검색을 실시하여 그 성능을 평가하였다. 선별된 16인은 다음과 같다.

이명박	박근혜	손학규	유시민	노무현
김정일	이재오	강재섭	홍준표	김대중
엄기영	최문순	황우여	오세훈	나경원
이광재				

제안 알고리즘을 연관검색에 활용하기 위해 그림 3의 초기 집합인 L_1은 최소한 3건 이상의 기사에서 언급된 인물들로 제한하였고, 이로부터 일대일 연관성을 갖는 집합인 L_2를 생성하였다. 일반적으로 연관검색 시스템에서는 특정 검색어에 대해서 직접적으로 연관된 검색어들을 순위별로 결과를 제공하므로 L_3 이상(크기가 3 이상인 후보 집합들)은 큰 의미가 없다. 생성된 L_2 집합 내에서 각 원소에 대한 연관성 정도는 지지도(support) 값을 활용하였다. 즉, 특정 검색어에 연관된 인물들에 대한 검색결과는, L_2로부터 검색어가 포함된 부분집합들을 추출하고 이들을 지지도를 기준으로 정렬하여 최상위 10인의 인물들을 추출하여 얻어진다. 예를 들어, 정치인 16인 중에서 '오세훈'과 연관된 인물을 찾기 위해 기존 Apriori와 Keyword_Apriori 알고리즘으로 검색된 상위 10위 결과는 표 1의 (a), (b)와 같다.

이와 같은 방식으로 인물 16인에 대해서 검색을 실시하여 그 결과를 비교하였다. 그런데 본 실험에서 실시한

두 가지 연관검색 방식의 정확도를 비교하기 위해서는 객관적으로 정확한 검색결과라고 판단할 수 있는 비교대상이 있어야한다. 이를 위해 본 논문에서는 해당 기간 동안 트위터에서 언급된 인물들 간의 연관정도 결과를 비교대상으로 활용하였다. 비록 트위터에서의 연관정도가 가장 정확한 결과라고는 할 수 없지만 트위터 기사들에서 연관성을 가진다면 대중적으로 그 연관성에 관심이 많다는 것에 대한 간접적인 표현이 될 수 있으므로 비교대상으로서 의미가 있다고 하겠다. 트위터에서의 인물에 대한 연관정도는 트렌드시크(http:// www.trendseek.co.kr)의 결과를 활용하였다. 트렌드시크는 SNS 데이터들을 분석하여 주요 검색어에 대한 동향을 분석하는 사이트이다. 예를 들어, 트위터에서 인물 16인 중 '오세훈'과 관련된 연관검색 결과는 표 1 (c)와 같다.

표 1. '오세훈'으로 검색된 연관인물 검색 결과
Table 1. Search results of associated people by the query 'Oh Sehoon'

(a) Apriori 검색결과		(b) Keyword_Apriori 검색결과		(c) 트위터 검색결과	
순위	검색결과	순위	검색결과	순위	검색결과
1	박근혜	1	남경필	1	김문수
2	남경필	2	박근혜	2	박근혜
3	손학규	3	정몽준	3	이명박
4	권영세	4	이재오	4	안상수
5	원희룡	5	손학규	5	이재오
6	이명박	6	김문수	6	손학규
7	나경원	7	유시민	7	정몽준
8	홍준표	8	이명박	8	유시민
9	유시민	9	권영세	9	남경필
10	박진	10	원희룡	10	원희룡

2. 성능평가 기준

본 실험에서 검색 결과를 비교하기 위한 측정치로서 DCG(Discounted Cumulative Gain)[14]를 활용하였다. DCG는 검색엔진의 효율성을 측정하는 도구로 가장 많이 쓰이는 기법중 하나이다. DCG는 검색결과로 나온 각 단어 - 일반적으로 문서 검색 엔진에서는 문서를 의미하나 본 논문에서는 연관 단어를 검색하므로 문서라는 용어 대신 단어라는 용어를 사용한다. - 에 대해서 검색어와의 실제 관련도에 따라 점수를 부여하는 방식을 사용한다. DCG에는 검색 결과의 순서(랭킹)보다 결과들의 관련도만을 정량화하여 계산하는 CG(Cumulative Gain) 방식이 있고, 검색결과의 랭킹이 중요한 요소로 활용하면서 정규화된 측정방식이 nDCG(Normalized Discounted Cumulative Gain)가 있다. 우선 CG는 검색 결과의 순서에 관계없이 각 결과 단어의 관련정도를 점수화 하여 그 값을 모두 합한 수치로 평가한다. 즉, 검색결과가 p개일 경우 CG는 다음의 수식으로 계산한다.

$$CG_p = \sum_{i=1}^{p} rel_i \quad (3)$$

여기서 rel_i는 검색 결과에서 i번째 위치한 단어의 관련도를 나타낸다. 반면에 DCG는 관련도가 높은 단어가 검색 결과에 우선적으로 랭킹된다는 사실에 입각해서 그렇지 않을 경우 감점을 부과하는 방식을 사용한다. 따라서 DCG는 다음의 수식으로 계산한다.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log 2(1+i)} \quad (4)$$

일반적으로 DCG는 위의 수식 그대로 사용하지 않고 0부터 1까지의 정규화된 수치인 nDCG로 표현하는데 수식은 다음과 같다.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (5)$$

$IDCG_p$는 관련도 값에 따라 정확하게 정렬되었다는 가정하의 DCG값을 의미한다. 따라서 nDCG는 1에 가까울수록 좋은 검색 결과를 나타낸다. 본 논문에서는 CG와 nDCG를 검색 결과의 평가에 모두 활용하였다. 단어 간의 관련도를 평가하기 위한 rel_i는 두 가지 방식을 사용하였다. 첫째는 검색 결과로 제공된 상위 10위의 각 인물에 대해서 비교대상인 트위터에서의 상위 10위에 존재하는 인물이면 1을 그렇지 않으면 0을 부여하였다. 즉, $rel_i \in \{0,1\}$이 되고 이를 절대관련도라 정의한다. 둘째는 검색 결과로 제공된 상위 10위의 각 인물에 대해서 트위터 검색결과에 상위 10위의 i번째 존재하면 $11-i$를 부여하고, 상위 10위에 존재하지 않으면 0을 부여한다. 즉, $rel_i \in \{0,1,...,10\}$의 값을 갖고 이를 상대관련도라 정의한다.

3. 성능평가 결과

성능평가는 표 2와 같이 CG와 nDCG 값을 측정하고 각각에 대해서 상대관련도와 절대관련도를 이용하여 계산하였다.

표 2. 평가 지표
Table 2. Evaluation metrics

방식 \ 관련도	상대관련도	절대관련도
CG	$CG_{10}^{constant}$	$CG_{10}^{relative}$
nDCG	$nDCG_{10}^{constant}$	$nDCG_{10}^{relative}$

성능 평가 결과는 그림 4~7에 제시되어 있다. 우선 그림 1은 $CG_{10}^{constant}$의 결과를 보여준다. 상위 10위을 검색했을 때 가장 이상적인 $CG_{10}^{constant}$의 값은 10이 된다. 이 그림에서 보는 바와 같이 각 16인의 검색어 중 절반인 8개의 검색어에서 Keyword_Apriori가 Apriori에 비해 더 우수한 성능을 보여주었고, 나머지 8개의 검색어는 동일한 결과를 나타냈다. 평균적으로는 14.3% 향상된 결과를 나타냈다. 그림 4는 $CG_{10}^{relative}$의 결과를 보여준다. 이상적인 $CG_{10}^{relative}$의 값은 1부터 10까지의 합인 55이다. 이 그림에서 보는 바와 같이 각 16인의 검색어중 12개의 검색어에서 Keyword_Apriori가 Apriori에 비해 더 우수한 성능을 보여주었고 나머지 4개의 검색어는 결과가 동일하였다. 평균적으로는 13.5% 향상된 결과를 나타냈다. 그림 4와 그림 5를 비교해보면 거의 유사한 결과를 나타낸다. 따라서 절대관련도와 상대관련도에 따른 서로간의 평가결과는 큰 차이가 없는 것을 확인할 수 있다.

그림 6은 $nDCG_{10}^{constant}$의 결과를 보여준다. 이 그림에서 보는 바와 같이 각 16인의 검색어중 13개의 검색어에서 Keyword_Apriori가 기존 Apriori에 비해 더 우수한 측정치를 보여주었고, 2개는 반대의 결과를 보였으며 나머지 1개의 동일한 결과를 나타냈다. 평균적으로는 9.1% 향상되었다. 마지막으로 그림 7은 $nDCG_{10}^{relative}$의 결과를 보여준다. 이 그림에서 보는 바와 같이 각 16인의 검색어 중에서 12개의 검색어에서 Keyword_Apriori가 Apriori에 비해 더 우수한 결과를 보여주었고, 2개는 오히려 성능이 떨어졌으며 나머지 2개는 동일하였다. 평균적으로는 8.5% 향상된 결과를 나타냈다. nDCG의 경우도 CG의 경우처럼 절대관련도와 상대관련도에 따른 서로간의 평가결과는 큰 차이가 없었다.

결론적으로 4가지 평가 항목에서 평균 8%~14%까지 검색 결과의 정확도가 향상되었음을 확인할 수 있었다. 그 이유는 기존 방식에 비해서 제안된 알고리즘이 기사의 특성을 이용하여 문장내에서 단어간의 관련도를 세밀하게 측정함으로써, 보다 정확한 검색이 가능하기 때문으로 분석할 수 있다. CG에 의한 평가 방식이 nDCG의 평가 방식보다 약간 우수하게 나타났다. CG는 연관된 단어가 얼마나 많이 상위권에 포함되었는가(본 실험에서는 상위10위)를나타내는 지표이고 nDCG는 상위권에서의 정확도가 얼마나 높은가를 판단하는 지표이다. 따라서 실험의 결과만으로 해석한다면 본 논문에서 제시한 방법은 상위에 랭크된 검색 결과의 순서보다는 상위권에 되도록 많은 연관된 단어를 포함하도록 검색하는 방법에 좀 더 적합하다고 해석할 수 있지만 그 차이가 크지 않으므로 큰 의미는 없다고 판단된다.

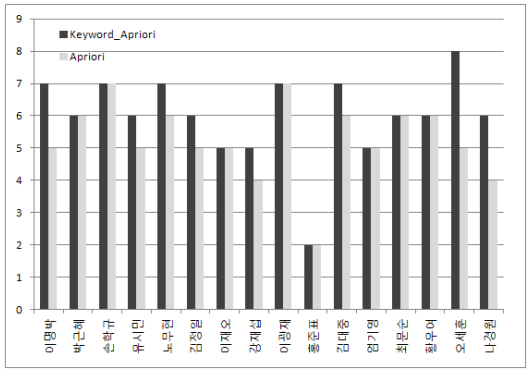


그림 4. $CG_{10}^{constant}$를 이용한 성능평가결과
Fig. 4. Empirical results using $CG_{10}^{constant}$

V. 결 론

본 논문은 연관규칙 마이닝 및 통계 처리를 바탕으로 뉴스 기사집합에 존재하는 연관된 키워드 집합을 도출하는 방안을 제안하였다. 연관 키워드 집합 추출은 연관검색 및 문서요약 시스템에 활용되며, 이는 사용자에게 검색 방향을 잡아주고 검색 효율을 높여준다는 측면에서 연구가치가 높다. 본 연구에서는 단어 개체간의 연관도의 정량화를 위해 TF-IDF 가중치 기법의 아이디어를 차용하였으며, 단어 개체의 문장내 공존성을 주요 인자로 삼았다. 연관 키워드 집합의 도출은 기존 Apriori 알고리즘을 변형한 Keyword_Apriori 알고리즘으로 실현된다. DCG 기반 성능평가를 통해 제안한 Keyword_Apriori 알고리즘이 기존 Apriori 알고리즘에 비해 최대 14% 우수

한 성능을 보였다. 향후 실험 데이터 집합을 인물 이외의 사회적 이슈가 되는 키워드들로 확대하여 TextMap[15]과 유사한 형태의 범용적 연관분석 시스템으로 발전시켜 나갈 예정이다.

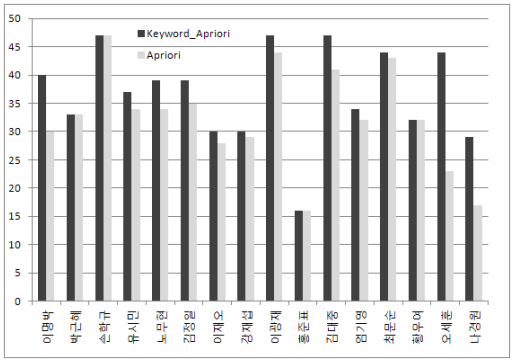


그림 5. $CG_{10}^{relative}$를 이용한 성능평가결과

Fig. 5. Empirical results using $CG_{10}^{relative}$

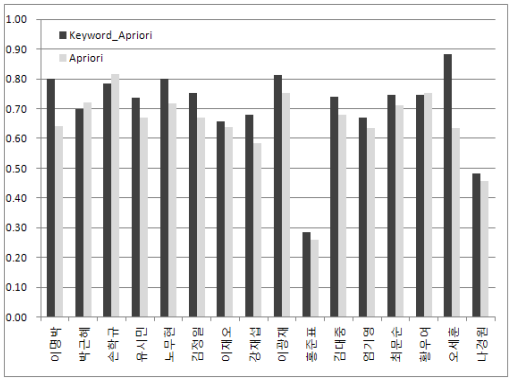


그림 6. $nDCG_{10}^{constant}$를 이용한 성능평가결과

Fig. 6. Empirical results using $nDCG_{10}^{constant}$

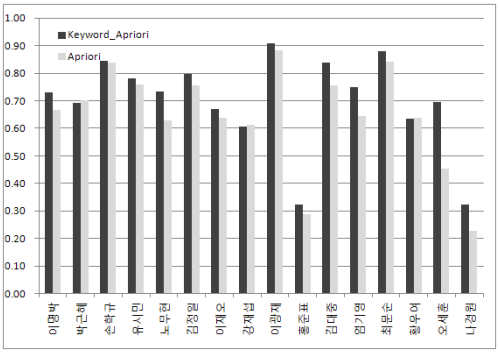


그림 7. $nDCG_{10}^{relative}$를 이용한 성능평가결과

Fig. 7. Empirical results using $nDCG_{10}^{relative}$

참 고 문 헌

[1] D. Ayers and A. Watt, Beginning RSS And Atom Programming, John Wiley & Sons Inc., 2005

[2] 이강표, 김두남, 김형주, “웹 2.0 환경에서의 태깅기술 동향”, 정보과학회지, 제25권 10호 pp. 36-42, 2007년 10월

[3] M. A. Hearst, D. Rosner, "Tag Clouds: Data Analysis Tool or Social Signaller?," Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), pp. 160~168, 2008

[4] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition), ACM, 2011

[5] http://lucene.apache.org/nutch/

[6] J. Hipp, U. Güntzer and G. Nakhaeizadeh, "Algorithms for Association Rule Mining: A General Survey and Comparison", ACM SIGKDD Explorations Newsletter Vol. 2, No. 1, 2000.

[7] S. E. Robertson, "The probability ranking principle in information retrieval", Journal of Documentation, Vol.33, pp.294-304, 1977.

[8] S. Lee, H. Kim, "News Keyword Extraction for Topic Tracking", Networked Computing and Advanced Information Management, Vol.2, pp.554-559, 2008

[9] R. Feldman, and I. Dagan, "KDT-Knowledge Discovery in Texts", Proceedings of the First International Conference on Knowledge Discovery (KDD), pp. 112 - 117, 1995.

[10] R. Feldman, and H. Hirsh, "Mining Associations in Text in The Presence of Background Knowledge", Knowledge Discovery and Data Mining, pp. 343 - 346, 1997.

[11] R. Feldman, I. Dagan, and H. Hirsh, "Mining Text Using Keyword Distributions", Journal of Intelligent Information Systems, Vol. 10, No. 3, pp. 281 - 300, 1998.

[12] H. Ahonen, O. Heinonen, M. Klemettinen, and I. Verkamo, "Applying Data Mining Techniques in Text Analysis", Technical Report C-1997-23, University of Helsinki, 1997..

[13] P. C. Wong, P. Whitney, and J. Thomas, "Visualizing Association Rules for Text Mining", IEEE Symposium on Information Visualization (INFOVIS), pp. 120 - 123, 1999.

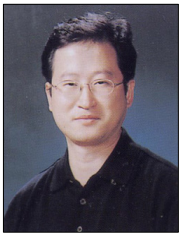
[14] http://en.wikipedia.org/wiki/Discounted_cumulative_gain

[15] http://www.textmap.com

※ 본 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것이며(과제번호: 2010-0025212), 또한 한성대학교 교내 연구비 지원과제임.

저자 소개

김 한 준(정회원)



- 1994년 서울대학교 계산통계학과 졸업 (이학사)
- 1996년 서울대학교 전산과학과 대학원 졸업 (이학석사)
- 2002년 서울대학교 컴퓨터공학부 대학원 졸업 (공학박사)
- 2002년~2002년 서울대학교 공과대학 Post-Doc
- 2002년~현재 서울시립대학교 전자전기컴퓨터공학부 부교수

<주관심분야> 데이터마이닝, 정보검색, 기계학습, 데이터베이스

장 재 영(정회원)

- 1992년: 서울대학교 계산통계학과 (이학사)
- 1994년: 서울대학교 계산통계학과 (이학석사)
- 1999년: 서울대학교 계산통계학과 (이학박사)
- 2000년~현재: 한성대학교 컴퓨터공학과 교수

<주관심분야> 데이터베이스, 정보검색, 데이터마이닝