효과적인 정보검색을 위한 개념망의 구축

주정은 구상회
고려대학교 디지털경영학과
{ kquilt, skoo }@korea.ac.kr

Construction of Concept Network Useful for Effective Information Retrieval

Joung Eun Ju, Sang Hoe Koo
Dept. of Digital Management, Korea University

요 약

본 연구에서는 정보 검색의 효과를 향상시키기 위한 방안으로 개념망을 제안한다. 개념망은 주어진 문서의 집합에서 제시된 주요 개념을 추출하고, 추출된 개념들 사이의 관련성을 분석하여, 관련성이 높은 개념 사이에는 링크를 설정함으로써 개념을 노드로 하는 네트워크를 구성한 것이다. 개념 추출과 링크 설정은 문서에 출현하는 명사의 출현 빈도를 근거로 하였다. 사용자가 정보검색을 위하여, 키워드를 입력하면, 본 시스템은 입력된 키워드를 중심으로 구축된 개념망을 제시한다. 사용자는 제시된 개념망을 조사함으로써, 자신이 입력한 단어가 검색하고자 하는 목표개념을 적절히 반영한 단어인지 확인할 수 있고, 새로운 검색어를 추가하거나 기존의 것을 수정함으로써 검색의 효과를 향상시킬 수 있다.

Ⅰ. 서론

월드와이드웹에서 제공되는 정보는 한곳에서 집중되어 관리되지 않고 웹에 연결된 수많은 컴퓨터에 분산되어 저장 관리된다. 또한 이렇게 분산되어 관리되는 정보는 성격, 내용, 표현 방식이 각 사이트별로 상이하게 다를 뿐 아니라, 시간의 흐름에 따라 역동적으로 변화한다. 따라서 웹에서 일반 사용자가 자신이 원하는 정보를 찾기란 쉬운 일이 아니다. 정보검색 서비스는 사용자들이 원하는 정보를 웹에서 효율적으로 찾아주는 기능으로서 현재 대다수의 포털 사이트에서 기본적으로 제공하고 있다[6].

현재 대부분의 정보검색 서비스는 질의어의 불리언 검색을 사용하여 웹페이지를 검색한 후, 검색된 페이지를 각자의 고유한 순위화 알고리즘에 의해 나열하여 사용자에게 제공한다. 그러나 이렇게 찾아진 검색결과가 검색 당사자의 의도를 정확히 반영하는 경우는 흔하지 않다. 이는 사용자가 자신의 검색의도를 불리언 질의어로 정확히 표현하기가 힘들고, 또한 검색 알고리즘과 순위화 알고리즘이 개별 사용자의 의도를 반영할 수 없기 때문이다[4].

본 연구에서 개발한 개념망(Concept Network)은 특정 분야의 주요 문서를 선정한 후, 이들 문서들의 내용을 분석하여, 문서에서 소개되는 주요 개념을 추출하고, 이들 개념 사이의 연관성을 분석하여, 연관성이 높은 개념들 사이에는 링크를 연결함으로써, 개념의 네트워크를 형성한 것이다.

정보검색에서 개념망을 사용할 경우, 사용자는 먼저 찾고자 하는 검색어를 입력한 후, 개념망을 통하여 찾아진 검색어와 관련된 기타 개념을 조사한 후, 검색어를 수정하거나 필요에 따라 추가하여 검색을 실행하게 된다.

개념망을 이용하면, 사용자는 개념망을 조사함으로써 자신이 입력한 단어가 검색하고자 하는 목표 개념을 적절히 반영한 단어인지 확인할 수 있으며, 필요에 따라 검색어를 수정하거나 추가할 수 있으므로, 검색의 효과성을 향상시킬 수 있다. 나아가 사용자가 미처 생각하지 못했지만 검색 목표와 관련있는 새로운 단어까지 찾아낼 수 있다.

본 논문은 다음과 같이 구성된다. 2절은 본 연구와 관련된 연구를 비교 분석하며, 3절은 본 연구의 결과인 개념을 추출하고 개념망을 구축하는 알고리즘을 설명하며, 4절은 본 연구를 실제 검색에 적용한 결과를 제시한다.

Ⅱ. 관련연구

본 절에서는 개념망을 다른 정보검색관련 연구와 비교 분석한다. 먼저 색인(indexing)은 어떤 문서를 다른 문서들로부터 구별할 수 있는 단어 또는 단어구를 추출하여 정보검색에 활용한 것으로 현재 대다수의 검색엔진이 사용하고 있는 기술이다. 색인의 경우, 사용자가 검색을 위해 선택한 검색어에 따라 쓸모없는 검색결과가 나올 수도 있다. 또한, 사용자가 원하는 적절한 결과를 얻기 위해 일반적으로 수 차례의 피드백 과정이 필요하다[4]. 반면에, 개념망은 사용자가 입력한

검색어와 유관한 개념들을 미리 제시해 사용자가 필요에 따라 새로운 검색어를 취하거나 또는 추가· 삭제하도록 함으로써, 검색의 효과를 향상시킬 수 있어 정보검색의 프리프로세서 역할을 한다.

시소러스(thesaurus)는 입력된 단어가 속하는 주제 분야의 중요 개념들을 대소관계, 동의관계, 동형이의관계, 상하관계 등의 관계성과 함께 제시하여 줌으로서, 검색의 효과를 향상시킨다는 면에서 본 연구가 제시하는 개념망과 유사하다. 그러나 시소러스는 구축에 소요되는 비용과 시간이 매우 크고, 수정 또한 용이하지 않다[3]. 개념망은 웹문서를 기반으로 자동으로 구축되기 때문에, 구축이 용이하며, 수시로 변화하는 웹 컨텐츠를 동적으로 반영할 수 있다.

AKN(Automated Keyword Network)은, 먼저 역색인 방식에 의해 자연어 색인을 추출한 후, 백과사전에 포함된 문장을 분석하여, 추출된 단어들 사이의 관계성을 찾아 이를 근거로 단어의 네트워크를 구성한 것이다[1]. AKN은 키워드를 추출하고 이들 간의 연관성을 조사하여 망을 형성한다는 점에서는 본 연구와 유사하나, 색인어가 포함된 문장의 구조적 정보(예를 들면, 표제어, 설명, 유의어 등)를 활용하여 색인어 사이의 연관성을 추정한다. 따라서, 백과사전과는 내용 구성이 구조적으로 다른 웹문서에 AKN을 적용하기는 힘들다.

Ⅲ. 개념망의 구축

1. 개념망의 정의

개념망(Concept Network)은 특정 분야의 주요 문서 집합을 선정한 후, 이들 문서들의 내용을 분석하여, 문서에서 소개되는 주요 단어를 추출하고, 이들 단어 사이의 연관성을 분석하여, 연관성이 높은 단어들 사이에는 링크를 연결함으로써, 단어의 네트워크를 형성한 것이다. 본 연구에서는 문서의 집합을 URL의 집합으로 표시하였다. URL의 집합은 디렉토리형 정보검색 사이트 등에서 카테고리 명을 조사하여 확보한다. 주요 개념은 이들 URL사이트의 웹문서 안에서 단어들이 출현하는 빈도를 근거로 산출한다. 개념망은 이들 주요개념사이의 연관성을 기반으로 구축하는데, 연관성은 모든 단어 쌍에 대하여 문서내에서의 공동 출현빈도를 계산하여 추정한다.

2. 주요 개념의 추출

개념 추출의 목적은 선택된 문서를 가장 잘 대표할 수 있는 중요 명사를 추출하는 것이다. 명사의 중요도는 전체 문서에서의 절대적 중요도와 개별 문서내에서의 상대적 중요도의 곱으로 계산한다. 절대적 중요도는 단어의 전체 문서에서의 총 출현 빈도로 계산되며, 상대적 중요도는 해당 단어의 개별 문서별 출현 빈도를 문서의 전체 단어 수로 나눈 것을, 모든 문서에 대하여 더한 것이다. 이렇게 계산된 명사의 중요도가 특정 임계치 이상의 값을 가질 경우 중요 개념으로 간주한다. 그리고 임계치 값은 실험을 통하여 추출된 개념을 조사함으로써 추정하였다. 아래는 개념을 추출하는 식을 표시한 것이다.

$$w_i = \text{절대가중치}_i \times \text{상대가중치}_i > \theta$$

$$= \text{워드 } i\text{의빈도} \times \sum_j \frac{\text{문서 } j\text{에서의 워드 } i \text{ 빈도}}{\text{문서 } j\text{의 모든 단어 빈도}} > \theta$$

$$= \sum_{j=1}^{n} wf_{ij} \times \sum_{j=1}^{n} \frac{wf_{ij}}{\sum_{k=1}^{m} wf_{kj}} > \theta$$

w_i: $word_i$의중요도

wf_{ij}: $word_i$의 doc_j에서의 $frequency$

m: $word$개수

n: doc(문서)개수

θ: 임계치

어떤 단어가 문서별로 동일하게 10회씩 출현했어도, 문서의 크기가 작은 문서에서 10회 출현한 단어가, 문서의 크기가 매우 큰 문서에서 10회 출현한 단어보다 중요하다고 할 수 있다. 따라서 상대적 중요도를 계산할 때, 단어의 출현빈도를 문서의 크기로 표준화(normalize)할 필요가 있다. 본 연구에서는 단어의 출현빈도를 문서의 크기로 나누어 표준화한다.

3. 개념망의 구축

개념망을 구축하는 목적은 사용자가 입력한 단어와 유관한 개념이나 단어를 사용자에게 제시함으로써, 사용자가 그의 검색의도를 보다 적절히 반영하는 키워드로 정보검색을 수행할 수 있도록 하는데 있다. 본 연구에서는 개념사이의 관계성을 단어쌍(word pair)의 동일 문서내 공동존재(coexistence) 정도에 의하여 결정한다. 즉, 어떤 두 단어가 같은 문서에서 매우 자주 출현한다면 두 단어는 관계성이 높은 개념으로 간주하며, 이 관계성이 특정 임계치 이상이면 두 개념 사이에 관계가 존재하는 것으로 간주한다. 이 같은 관계성을 앞 절에서 찾은 모든 단어의 쌍에 대하여 조사하여, 임계치 이상인 경우 두 단어 사이에 링크를 설정하여 개념망을 구축한다.

때로는 두 단어의 절대 출현 빈도가 지나치게 커서, 두 단어 사이의 관계성 때문이 아니라, 절대출현빈도에 비례하여 공동출현 빈도가 커지는 경우가 있을 것이다. 이런 경우를 배제하기 위하여 공동존재 빈도를 표준화(normailize)할 필요가 있다. 본 연구에서는 두 단어의 공동출현 빈도를 두 단어의 총 출현빈도로 나누어 표준화한다. 단어 i와 단어 j의 연관성 정도를 나타내는 식은 아래와 같다. 식에서 임계치 값은 실험을 통하여 추출된 개념을 조사함으로써 추정하였다.

$$R(w_i, w_j) = \frac{\text{단어}i\text{와 단어}j\text{의 문서내 공동출현빈도}}{\text{단어}i\text{와 단어}j\text{의 전체출현빈도}} = \frac{\sum_{k=1}^{n} MIN(wf_{ik}, wf_{jk})}{\sum_{k=1}^{n} wf_{ik} + \sum_{k=1}^{n} wf_{jk}}$$

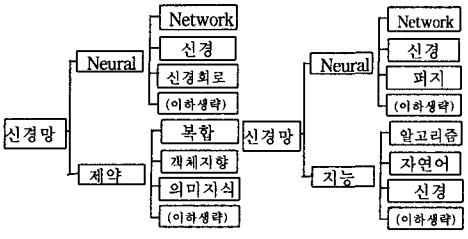
$R(w_i, w_j)$: $word_i$와 $word_j$사이의관계성
wf_{ij}: $word_i$의 doc_j에서의 $frequency$
θ: 임계치

Ⅳ. 실험

본 절에서는, 본 연구가 제시한 개념망이 실제 정보검색에 적용할 때 얼마나 효과적인지를 실험을 통해 알아보았다. KTSET문서[8]는 태그 정보와 컨텐츠 정보로 이루어져 있기 때문에, 일반적인 웹 문서와 구조적으로 유사하며, 정보검색의 성능을 판단하기 위해 가장 널리 사용되는 표준화된 문서이다. 따라서 본 연구에서는 KTSET를 사용한다. 본 프로그램은 VISUAL BASIC으로 작성되었으며, 명사 추출하는 부분은 HAM5.0[3,7]을 이용하였다.

KTSET문서를 일정 수만큼 수집하여 한 그룹으로 URL 그룹을 만들고, 이 그룹안에 있는 문서의 태그를 제거한다. 다음으로 HAM으로 개념추출 대상이 될 명사들을 추출하고 이들 명사 중에서 URL 그룹을 대표할 수 있는 개념들을 선정하게 된다. 임계치를 10, 25, 50, 100 등으로 나누어서 비교해 본 결과 임계치가 100인 경우 영역전문가가 판단하기에 가장 적절한 개념들이 추출되었다.

이런 과정을 거쳐 추출된 개념을 대상으로 단어쌍을 만들어서 임계치 이하인 단어쌍들은 개념망을 구축하게 된다. 본 실험에서는 0부터 0.5 사이의 숫자를 이용하여 실험을 해 보았다. 그 결과, 0.25인 경우 영역전문가가 가장 선호하는 결과를 생성되었다.



< 그림1 시스템>

<그림2 전문가 그룹>

<그림 1>은 단어 "신경망"에 대해 본 연구에서 구현된 시스템이 찾아낸 개념망을 트리형식으로 나타낸 것이다. 위 그림의 트리는 개념망을 사용자가 보기 편하도록 트리형태로 나타낸 것으로, 개념망을 넓이 우선으로 탐색할 경우 탐색 순서를 트리로 표현한 것이다. <그림 2>는 본 시스템이 아닌, 영역전문가들이 문서의 내용을 읽어본 후, 수작업으로 작성한 개념망이다. 위 그림에서 볼 수 있듯이, "신경망"이라는 키워드가 제시된 경우, 본 시스템과 영역전문가가 작성한 개념망이 매우 유사함을 확인할 수 있다.

Ⅴ. 결론

개념망은 주요개념을 추출하고 이 개념들을 망으로 형성한 것이다. 주요개념은 문서상에서의 출현빈도를 근거로 절대적 중요도와 상대적 중요도를 계산하여, 이 중요도가 특정 임계치 이상인 명사를 대상으로 선정하였다. 개념망은 이들 주요 개념을 대상으로 공동출현 빈도를 기반으로 개념 사이의 관련성을 추정하여, 이 관련성이 특정 임계치 이상인 경우 링크를 설정함으로써 구축된다.

정보검색 시 개념망을 사용하면, 사용자가 검색을 위해 입력한 단어의 개념망을 조사할 수 있다. 사용자는 이 개념망을 조사함으로써, 자신이 검색하고자 하는 목표 개념과의 일치 여부를 쉽게 파악하고 더 나아가 사용자가 미처 생각하지 못했지만 연관있는 개념까지 정보검색에 활용할 수 있다. 본 연구에서 개발한 개념망을 적절히 활용하면, 불린 방식 검색의 효과를 높일 수 있을 것으로 기대된다.

[참고문헌]

[1] 김정세 외 1인 , 2000, 정보검색에서의 어의 중의성 해소를 위한 자동키워드망의 이용, 한국 정보 처리학회 논문지 제7권 제 12호
[2] 김태수 외 1인, 1999, WordNet과 시소러스, 언어정보의 탐구, pp232-269
[3] 임형근 외 1인 , 2001, 색인어 연관성을 이용한 의료정보문서 분류에 관한 연구, 한국정보처리학회 논문지B 제 8-B권 제 5호
[4] 정영미, 1993, 정보검색론, 구미무역(주)출판부
[5] 최재훈 외 2인, 2000, 객체기반 시소러스 시스템의 설계 및 구현, 정보과학회 논문지 제 27권 제 1호
[6]Baeza-yates 외 1인, 1999,최신 정보검색론, 홍릉과학출판사
[7]http://nlp.kookmin.ac.kr/, 국민대학교 자연언어 정보검색 연구실
[8]http://green.skhu.ac.kr/~skhuir/, KTSET문서