지식 정보를 이용한 웹 문서 순위 결정 방법

* 이용현 변영태 구연건
홍익대학교 전자공학과 홍익대학교 컴퓨터 공학과 홍익대학교 전자공학과

A Web Document Ranking Method Using Knowledge Information

Rhi, Yong-Hyun Byun, Young-Tae Koo, Yeon-Geon
Hongik University Hongik University Hongik University

요약

디양하고 방대한 양의 정보가 산재해 있는 웹의 정보 소스들 중 특정 도메인에 해당하는 문서를 판단하는 방법과 효과적인 검색을 지원하기 위한 웹 문서 순위 결정 방법을 제안한다 웹에 산재되어 있는 정보들은 특정 도메인에 적합한 문서들만을 저장하기 위해 특정 지식 베이스를 기반으로 하는 지식 정보를 이용하여 문서의 평가 작업을 한다 또한 효과적인 검색을 지원하기 위한 웹 문서 순위 결정 방법으로 두 단계의 처리를 거친다 첫째, 사용자로부터 주어진 질의에 대해 그 질의를 이루고 있는 각 단어들에 대한 상대적인 중요도를 반영한 P-norm 모델을 적용하여 문서의 순위를 결정한다, 두 번째 단계로 관련 문서들간의 유용한 링크 정보를 이용한 재조정 작업에 의해 문서의 순위를 결정하는 방법을 제시한다. 본 논문에서 제시한 방법에 의해 적절한 문서의 판단과 더 나은 정보의 제공을 수행할 수 있는 효과적인 정보 검색 기능을 가진다 정보 탐색자에게 정확하고 효과적인 정보를 제공할 수 있게 되었고, 일반적인 웹 검색 방법과의 비교에 의해 검색 효율이 향상됨을 보인다.

1. 서론

컴퓨터와 네트워크의 급속한 발달로 현대 사회는 정보화 사회로 급변하고 있다 이와 더불어 최근에는 인터넷의 급속한 보급으로 정보의 제공자와 정보의 사용자가 급증하고 있다. 이와 같이 웹을 통한 정보의 접근은 일반화되어 기고 있지만 무수히 쏟아지는 엄청난 양의 정보와 그 디양성은 효율적인 정보 검색의 커다란 장애 요인이 되고 있다

본 논문은 이런 환경에서 웹을 통해 모아진 정보들 중에 특정 도메인에 적합한 문서를 효율적으로 평가하는 방법과 웹 문서를 대상으로 효과적인 검색을 치리히기 위해서 순위 결정 방법을 제안헌다 웹 상의 문서는 특징 도메인에 대한 문서의 적합성 여부를 판단하기 위하여 특정 영역의 지식 기반의 지식 정보를 이용한 문서의 평가 작업을 치리하게 된다 또한 문서의 정합(matching) 과정에서 주어진 불리언 질의의 탐색어들에 대한 상대적인 중요도의 처리와 문서 구조의 유용한 링크 정보를 이용하여 문서의 순위를 결정히는 방법을 제공함으로서 향상된 검색 신뢰도를 제공하게 된다

디양힌 형태의 웹 문서를 색인하기 위해서는 색인어의 정보를 기반으로 정보 추출이 이루어지고, 이런 색인어들의 정보를 제공하는 방법으로 시소러스[1]나 상호정보[2] 지식베이스를 이용하는 방법이 연구되고 있다 이런 방법들은 문서의 내용을 알아내는 지식 정보의 의미로도 사용된다. 추출된 색인어들에 대한 문서상의 중요도에 관한 대표적인 연구로는 각 색인어들에 대한 여러 가지 가중치 요소에 대해 연구되고 있다 [3] 그리고 관련 문서들의 순위를 설하기 위해서는 각 문서간의 유사도를 계산하여 문시의 검색 여부니 순위를 결정하는 알고리즘들이 사용되는데, 그 중 대표적인 검색 모델로는 Vector space model 과 Probabilistic model[4], Fuzzy set model[5], Extended boolean model[6] 등에 대한 연구기 이루어지고 있다 특히, 웹 기반의 문서의 구조에서 링크 정보를 이용한 순위 처리 방법에 의한 순위 결정 방법[7]이나, 관련 문서 집합에서의 링크 정보로부터 순위를 재처리 방법[8]에 대한 연구가 진행되고 있다.

2장에는 문서의 적합성을 판단히는 방법에 대해 알아보고, 3장에서는 순위 결정 방법에 대해 설명한다. 그리고 4장에서는 실험을 통해 분석 내용을 기술하고, 마지막으로 결론을 내리도록 한다

2. 문서의 적합성 판단

색인은 문서와 질의간의 유사도를 계산하기 위한 정합(matching) 과정에서 사용할 수 있도록 문서에서 필요한 정보를 추출하는 과정이다. 이런 색인 과정에서 방대한 양의 정보를 가진 웹 문서들 중 특정 도메인에 해당하는 정보 문서만을 수집하고 문서의 평가를 거쳐 현 도메인에 적절한 문서라고 판단하는 것은 중요한 일이다 이런 문서의 적합성 판단은 시스템의 성능에 아주 중대한 영향을 미친다 본 논문에서는 특정 도메인으로 영역을 제한하기 위하여 동물에 관련된 내용을 기반으로 구축된 지식 베이스[9]를 사용하여 지식 정보를 획득한다 이 지식 정보와 문서는 평가 함수에 의해 그 문서의

본 연구는 과학 재단(과제 번호 973010301)의 지원을 받았음

적합치를 계산하게 된다. 이를 바탕으로 각 문서들은 특정 영역의 지식 검보에 대한 평가 값을 부여받게 된다 이는 문서의 적합성 여부를 판단하는 값으로 작용하게 된다

2.1 지식 정보의 획득

동물 관련 지식 베이스[9]는 지식을 효율적으로 표현하는 방법으로 구조적인 표현 방법을 사용한다 이 지식 베이스는 동물 분야에 대한 전문적인 지식을 갖는 도메인 전문가(domain expert)의 역할을 한다 이 지식베이스를 이루고 있는 계층 구조의 각 객체(object)들과 그들의 속성(slot) 그리고 속성에 관련된 값(value)을 가지고 있다

이런 지식 베이스와 연계하여 계층 구조를 이루고 있는 객체(object)와 그 동의어들을 동물 도메인의 고유한 특성으로 간주하여 계층구조 관련 색인어들의 집합으로 처리한다 또한 객체의 속성이나 속성 값들은 동물에 관련된 고유 특성의 의미는 약하지만 중요한 정보라는 의미에서 속성 관련 색인어들의 집합으로 처리한다 이런 각각의 색인어 집합들은 지식베이스에 있는 지식을 정보화 한 형태로 문서의 적합성을 판단하는 데 기준이 된다.

2 2 문서의 표현

문서의 적합성을 판단하기 위해서는 먼저 문서를 분석하여 색인어의 형태로 변환하는 과정이 필요하다 이의 처리 과정은 입력 문서들의 구성 문장들을 가져와 형태소 분석과 태깅(tagging) 과정을 거치고 난 뒤 색인어를 추출하게된다 이렇게 선정된 색인어들은 색인어로서 가치가 없는 불용어를 처리하는 stopping 과정을 거친다 남은 색인어들은 단어들의 일관성을 위하여 어간 처리 작업인 stemming 과정을 처리한다 이런 작업을 마친 후 남은 색인어들로 posting file을 만들게 된다 이 posting file은 색인어, 출현 빈도수(term frequency)의 쌍으로 구축되어 지게 된다 이 posting file은 문서의 적합성 여부의 판정에 사용하게 된다 여기서, 출현 빈도 수(tf)는 문서에 나타난 색인어의 빈도 수이다.

2 3 평가 함수

평가 함수는 그 문서의 posting file의 각 색인어와 비교 대상이 되는 평가 단어의 집합을 비교해 가면서 문서의 적합성 평가 값인 RE(Relevance Evaluation)를 구해낸다 문서의 posting file은 각 색인어(t)와 그 색인어의 출현 빈도 수인 중요도(w)로 구성되어 있다 즉, 문서는 $D=\{(t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n)\}$의 형태일 때, 평가 함수 식은 다음과 같다

$$RE(D) = \left\{ \sum_{i=0}^{n} re(t_i) \right\} / \ n \qquad (1)$$

$$re(t_i) = \begin{cases} \alpha * w_i & \text{if } t_i \in H \\ \beta * w_i & \text{if } t_i \in P \\ 0 & otherwise \end{cases} \qquad (2)$$

여기서 n은 문서의 전체 색인어의 개수이고, H는 계층 구조 관련 색인어 집합에 속하는 색인어들의 집합이고 P는 속성 관련 색인어 집합에 속하는 색인어들의 집합으로 지식 정보의 획득 과정에서 이미 생성된 집합이다. α, β는 각 색인어 집합에 나타난 경우의 중요도 값이다 ($0\le\alpha\le1, \alpha>\beta, \ \beta\ge0$)

모든 문서는 posting file을 만드는 작업과 이 평가 함수에 의해 자신의 평가 값인 RE를 가진다 이 값은 해당 문서에 대해 H, P가 가지고 있는 특정 도메인과 비교해 얼마나 도메인에 적합한 지를 판정해 주는 값이다.

3. 문서의 순위 결정

주어진 불리언 질의의 각 탐색어가 나타나는 문서들을 탐색하여 찾아내게 된다 이렇게 찾아낸 문서들은 질의에 관련된 문서의 집합으로 간주한다 문서 집합에 있는 문서들은 질의와 각 문서들간의 유사도 계산에 의해 각 문서의 중요도를 계산할 수 있고 이를 바탕으로 문서의 순위를 결정하게 된다

본 논문에서는 관련 문서의 순위를 개선하여 더 나은 정보의 제공을 위하여 두 가지 순위 결정 방법을 제안한다 첫째로, 질의의 탐색어들간의 상대적 중요도를 감안한 P-norm 모델의 이용으로 순위 결정을 처리한다 두 번째로는 추출된 관련 문서들간의 링크 정보를 이용하여 문서의 제 순위 결정을 처리하여 순위가 재조정된 관련 문서들을 더 돌려줌으로서 검색의 신뢰도를 높여준다

3.1 질의의 상대적 중요도에 따른 순위 결정 방법

질의와 문서사이의 유사도 값을 계산하는 P-norm 모델에서 질의의 각 탐색어는 자신의 중요도 값으로 유사도 연산에 참여한다 탐색어의 중요도를 구하는 방법은 존재 유무에 따라 0 혹은 1을 부여하는 weight 적용, 또는 출현 빈도 수(term frequency)의 적용[3] 등 여러 가지 방법을 가지고 있지만 일반적으로 질의에 속하는 탐색어는 중요도로 같은 값을 취하게 되는 경우가 많고 이런 방법은 질의의 각 탐색어들에 대한 상대적인 중요도를 부여하는데 영향을 미치지 못한다

본 논문에서는 이런 질의의 각 탐색어들간의 상대적인 중요도를 부여하기 위하여 퍼지 집합 모델을 근거로 각 탐색어들간의 중요도 관계를 제시해본다 연산의 순서를 오른쪽에서 왼쪽으로 진행한 경우 질의의 마지막 탐색어의 중요도를 1이라 할 때, 처음 탐색어의 중요도는 $2^{n-2}(n\ge2)$로 해 본다 결론적으로 인접한 탐색어사이의 중요도는 차음 계산된 두 개의 탐색어만 같을 뿐, 나머지 탐색어사이에는 1/2의 차이가 발생되는 것이다 이는 질의 $Q=\{(q_1,w_1),(q_2,w_2),\cdots,(q_n,w_n)\}$가 주어졌을 때 각 탐색어의 순서에 따른 상대적 중요도는 다음과 같이 나타난다

$$w_1 : w_2 : \cdots : w_n = 2^{n-2} : 2^{n-3} : \cdots \ 2^{n-n} : 2^{n-n} \qquad (3)$$

식 3에 의해서 질의 중요도를 감안하면 질의가 (t_1,t_2,t_3)인 경우 각 중요도는 (2,1,1)의 형태가 된다 질의는 먼저 나온 탐색어의 중요도와 정확성이 더 높다고 평가하여, P-norm 모델에 의해 문서와 질의간의 유사도 값을 계산할 때 질의의 각 탐색어의 가중치에 식 3을 적용함으로서 탐색어간의 중요도를 포함한 유사도 값을 계산할 수 있다.

3 2 검색된 문서의 링크 정보를 이용한 순위 재조정

웹 문서의 구조적인 특징 중 하나가 문서들간에 링크 정보가 구축되어 있다 즉, 하이퍼미디어에 대한 검색에서 링크 정보는 문서의 신뢰도를 결정하는 유용한 정보이다 이런 링크는 여러 형태로 구분할 수 있다. 즉, 링크의 방향성(incoming link outgoing link)과

링크의 관련성(앵커 단어와 키워드와의 정합)에 그 의미를 둘 수 있다. 본 논문에서는 탐색어의 상대적인 중요도를 포함한 유사도의 계산 후 관련 문서의 링크 정보를 다음과 같이 고려한 순위 재조정 단계를 거친다

$$Sim(D_i) = Sim(D_i) + \sum_{j=1, j\neq i}^{N} \alpha Li_{i,j} \cdot Sim(D_j) + \sum_{j=1, j\neq i}^{N} \beta Lo_{i,j} \cdot Sim(D_j) \quad (4)$$

N 추출된 관련 문서의 총 개수

$Sim(D_i)$: 관련 문서 중 i번째 문서의 유사도 값

$Li_{i,j}$ 문서 j에서 문서 i로의 incoming 링크의 발생 유무

$Lo_{i,j}$ 문서 i에서 문서 j로의 outgoing 링크의 발생 유무

α, β는 각 link의 가중치로 $0<\alpha<1$, $0<\beta<1$이다.

이와 같은 링크 정보를 이용한 유사도의 조정은 문서와 링크 되어 있는 문서의 개수뿐만 아니라 링크 되어 있는 문서의 유사도 값을 참조하여 관련된 문서 사이의 종속 관계를 표현할 수 있다.

4. 실험 및 분석

본 논문은 window NT환경에서 DBMS로 MS/SQL server를 사용하고 구현은 Visual C++을 사용해서 구현하였다. 특정 도메인에 적합한 문서를 추려내는 평가함수는 동물에 관련된 내용을 기반으로 평가 함수식에 의해 추출된 문서를 평가한 결과를 표 1에 보여 준다

표 1에의 실험 결과에서 "polar bear"라는 키워드를 검색 엔진을 통해 검색된 200개의 문서 중 평가 값이 0.9이상인 문서를 추출한 결과 전체 문서 중 약 29%의 문서가 적합 판정을 받았고, 이를 조사해 본 결과 약 79% 가량의 문서가 생물의 관점에서 polar bear에 관련된 문서임이 밝혀졌다

키워드	polar bear	비고
검색 엔진에서 검색된 문서	200개	
적합하다고 판정된 문서	57개	29%
적합 문서 중 생물 관련 문서	45개	79%

[표 1] 실험 결과

불리언 질의가 주어졌을 때 단지 그 질의에 의한 키워드 매칭에 의해 추출된 문서와 식 3에 나타난 질의의 상대적 중요도를 적용한 경우의 비교를 했다. 그 결과 사용자의 의도와 질의의 각 탐색어의 순서와 잘 나타난 경우에 처음 질의어에 대한 문서들이 높은 순위에 위치되어 있음을 확인했다

또한 1차 검색된 문서와 식 4에 의해서 유사도 값을 재조정한 경우의 문서를 비교한다 식 4의 α, β는 0.5로 같이 적용했다. 그 결과 문서의 순위의 변화가 나타나고, 전문가에 의해 더 나은 순위임을 확인했다

5. 결론 및 향후 과제

본 논문에서는 웹 문서의 적합성에 대한 판단의 방법으로 특정 영역의 지식 기반을 사용한 지식 정보를 이용하여 문서의 평가 값을 계산할 수 있다. 이는 현 도메인 지식 정보외의 각 문서의 상관 관계를 명시할 수 있는 방법으로 문서의 내용에 대한 적절한 판단이 이루어 질 수 있다

또한, 주어진 불리언 질의에 대한 문서의 검색에서 각 탐색어에 대한 상대적인 중요도를 감안한 순위 결정과 함께 2단계로 추출된 문서간의 incoming link와 outgoing link 관계를 가지고 있는 문서들간의 유사도의 값을 참조할 수 있도록 함으로서 검색의 신뢰도를 향상시키게 되었으며 이러한 문서의 순위 결정 방법을 실험을 통해 나타난 결과로 비교, 분석하였다

향후 본 논문에서 나타난 문서의 적합성 판단 방법과 순위 결정 방법을 포함한 웹 정보 검색 시스템의 개발과 지식 베이스를 이용한 질의 처리기와 에이전트의 개념을 포함한 지능형 정보 에이전트의 개발에 대한 연구를 지속해 나갈 것이다

참 고 문 헌

[1] Roy Rade & Ellen Bicknell, "Ranking Documents with a Thesaurus," Journal of the American Society for Information Science, 1989

[2] H K Kang, K.S. Choi, "TWO-LEVEL DOCUMENT RANKING USING MUTUAL INFORMATION IN NATURAL LANGUAGE INFORMATION RETRIEVAL," Information Processing & Management, Vol 33, NO 3, pp289-306, 1997

[3] G Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, Vol 24, No 5, pp 513-523, 1988.

[4] D Harman, "Ranking Algorithms," Information retrieval Data Structure & Algorithms pp 363 - 392

[5] 이준호, 김명호, 이윤준, "긍정적 보상 연산자를 이용한 피지 집합 모델의 개선", 한국 정보 과학회 논문지, 1993

[6] E A. Fox, S Betrabet, M Koushik "Extended Boolean Models," Information retrieval Data Structure & Algorithms pp 393 - 406

[7] B Yuwono, D L Lee, "Search and Ranking Algorithms for Locating Resources on the WWW," Proceedings of the 12th International Conference on Data Engineering, pp164-171, 1996

[8] J. Savoy, "A LEARNING SCHEME FOR INFORMATION RETRIEVAL IN HYPERTEXT," Information Processing & management, Vol 30 No 4 pp.515-533, 1994

[9] 오정옥, 변영태, "정보에이전트를 위한 지식 기반(동물) 질의 처리 시스템," 1998