94-a-5

용어의존도 모형에 의한 문헌검색가의 결정

the Term Dependence Models in a decision of a Retrieval Status Value

이 효 숙 이화여자대학교 도서관학과 강사

Lee Hyo Sook Ewha Woman's University, Lecturer

the Term Dependence Models has been examined in vew of the theoretical aspects. For the feasibility of these models in an operational system, the parameter estimation problem and other alternative solution has been included.

1. 서언

정보검색에서 용어 간의 의존도를 고려할 때, 의존도는 용어의 분포패턴에 의한 상관관계로서 측정된 의존도를 의미한다(van Rijsbergen, C. J., 1979). 용어 간의 의존도는 검색될 문헌의 적합성 추정에 있어서 중요하며, 결과적으로 검색성능에 영향을 준다.

일반적으로 확률검색에서 문헌에 대한 결정규칙을 적용하여 용어 간의 의존도를 이용하고 있다. 용어 의존도 모형들에 대해 다음에서 살펴보고, 실제 검색에서 적용할 때에 고려하여야 할 주요문제들을 논하고자 한다.

2. 용어의존도 모형

2.1 트리 모형 (the Tree Dependence Model)

트리 모형에서는 용어 간의 의존도가 하나의 트리를 구성하는 것으로 보고 각 정점(vertex)과 간선(edge)으로 나타낸다 (van Rijsbergen, 1977).

공식화하면 G = (V, E)로서 정점과 방향을 갖는 간선들로 이루어진 집합으로 <그림-1>과 같이 표현된다.

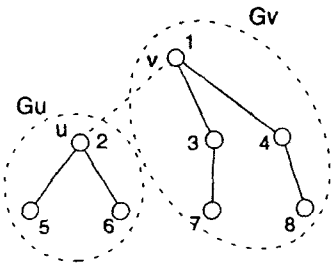

그림 1. 그래프 G 를 이루는 부분그래프 G_v, G_u

트리구조를 갖는 용어의 확률분포는 식 (1)과 같다 (Yu, C.T., Buckley, C., Lam, K. & Salton, G., 1983).

$$f(\underline{x}\ ;\ G) = p(v)\ \left[\prod_{E} P(a/b)\right]$$

$$f(\underline{x}\ ;\ G) = \frac{p(u,v)}{p(u)p(v)}\, f(x_v;G_v)\ f(x_u;G_w)\cdots(1)$$

x_v와 v_u는 부모노드 v를 갖는 정점의 집합 v_v와 부모노드 u를 갖는 정점의 집합 v_u에 한정하는 변수이다. $\underline{x}$는 용어 $x_1, x_2, \ldots x_n$으로서 n 개의 용어벡터를 나타낸다.

트리 모형은 다음의 속성을 갖는다.

첫째, 트리 모형의 그래프 상에서 트리 의존도는 선택한 루트나 간선의 방향에 관련없이 동일한 확률분포를 적용한다.

둘째, 정점 i와 정점 j 간에 정점 t가 있을 경우에 i와 j의 상관관계는 i와 k, k와 j의 상관관계의 곱이 된다.

2.2 BLE 모형 (Bahadur-Lazarsfeld Expansion Model)

BLE 모형에서는 0과 1로 표현되는 q 차원의 문헌벡터에 대해 가능한 한 정확하게 추정하기 위해서 용어 x_1,.... 용어 x_q에서 1의 값을 갖는 경우에 대해 식 (2) 로서 문헌에서의 용어의 확률분포를 산출한다.

$$p(x) = W_o(x)\{1 + \sum_{i<j} \rho_{ij}\, \frac{(x_i-p_j)(x_j-p_j)}{\sqrt{p_ip_j(1-p_j)(1-p_j)}} + \cdots + \rho_{12\cdots q}\, \frac{(x_1-p_1)(x_q-p_q)}{\sqrt{p_1\cdots p_q(1-p_1)(1-p_q)}} \cdots (2)$$

W_o는 용어가 독립적으로 출현하는 경우의 확률 분포로서 $\prod_{t=1}^{q} P_t^{x_t}(1-p_t)^{1-x_t}$이고, ρ_{ij}는 용어 i와 용어 j 간의 상관계수 값이다.

BLE 모형에서는 문헌집단에서 용어들이 독립적으로 출현한 경우와 용어가 독립적으로 출현하지 않은 경우에 대해 모든 가능한 용어 간의 관계를 고려하여 용어의 확률분포값 P(x)를 계산한다. 질문 (Q)에 대한 문헌의 검색가는 $\sum P(x)$가 된다.

2.3 확장트리 모형 (Generalized Dependence Model)

확장트리 모형은 트리 모형에서 세 용어 이상의 용어 간 의존관계를 검색에서 적용하도록 제시된 모형으로서 실제적으로 세 용어 간에 포함된 의존도로서 충분한 것으로 보고되었다 (Salton, G. 1989).

정점 u,v,w와 간선(u,v,w)를 구성하는 그래프에서 이들 용어의 확률분포 값은 식 (3)이 된다.

$$f(x\ ;\ G) = \frac{p(u,v,w)}{p(u)p(v)p(w)} \cdot f(x_u;G_u)\ f(x_v\ ;\ G_v)\ f(x_w\ ;\ G_w) \quad \cdots\cdots (3)$$

확장트리 모형은 검색에서 그래프 G^i와 그래프 G^{i+1}를 사용할 때 세 용어 u,v,w에 있어 세 용어 간의 관계 (u,v,w)를 사용할 때와 두 용어

간의 관계 (u,w) 및 (v,w) 와는 차이가 있다는 데에 촛점을 둔다. 이 모형에서는 식 (4)의 값이 최대가 되도록 의존도 트리에 간선을 선택적으로 추가함으로서 용어의 확률분포에 대한 근사화를 개선하도록 한다.

$$w = \sum_{u,v,w} p(u,v,w) \log \frac{p(u,v,w)}{p(w)p(u/w)p(v/w)} \quad \cdots(4)$$

2.4 질문이용에 의한 모형

질문구조를 이용한 의존도 모형은 검색시에 불 논리에 의한 질문에서 논리연산자 AND와 OR 관계를 갖는 용어들을 조사하여 이 용어들로 각 단계의 노드로 구성된 용어그룹을 형성한다. 그리고 문헌의 검색가 산출에서 이 용어 간의 관계를 사용한다.

문헌에 대한 최적의 순위화 함수는 식 (5)를 사용한다.

$$\log \frac{p(x)}{Q(x)} \simeq \log \frac{p'(x)}{Q'(x)} + A \quad \cdots\cdots\cdots\cdots (5)$$

식 (5)에서 $\log \frac{p'(x)}{Q'(x)}$는 용어 간의 독립성 가정에 의한 문헌 검색가로서 최종적인 문헌 검색가는 A 부분이 추가됨으로서 수정된다. A 부분은 용어간 의 의존도를 갖는 문헌에 대한 것으로 실제 계산은 용어의 확률분포 값과 용어간 상관계수 값을 사용한다.

3. 검색에서의 적용

용어 의존도 모형들을 정보검색에서 적용하기 위해서는 용어 간 의존도의 측정과 의존도 트리를 구성하여야 한다. 그리고 문헌검색가를 산출하기 위해 결정함수를 사용한다.

3.1 용어 간 의존도

용어 i와 용어 j 간의 의존도에 대한 측정척도로 EMIM(the expected information measure)이 제시되었다(van Rijsbergen, 1979). EMIM의 사용에서는 각 용어의 출현빈도와 각 용어의 동시출현빈도를 사용하여 식 (6)의 값을 산출한다.

$$I(x_i, x_j) = \sum_{x_i x_j} p(x_i, x_j) \log \frac{p(x_i,x_j)}{p(x_i)p(x_j)} \quad \cdots\cdots (6)$$

두 용어가 문헌에서의 출현이 독립적인 경우에는 $p(x_i)\ p(x_j) = p(x_i, x_j)$이므로 $I(x_i, x_j)$의 값은 0이 된다. EMIM 외에도 마론과 쿤스(Maron M.E. & Kuhns J.L.), 아이비(Ivie)등이 제시한 용어간 연관척도가 사용될 수 있다. (Maron & Kuhns, 1960; Ivie, 1966)

3.2. 의존도 트리의 구성

용어 쌍간의 의존도 수준에 따라 MST (Maximum Spanning Tree)를 생성하기 위해서 개발된 알고리즘들이 있으며 가장 효과적인 것으로는 휘트니(Whitney)의 알고리즘이 소개되고 있다(Frakes & Baeza-Yater, R., 1992). FORTRAN언어로 구성된 휘트니 알고리즘에서는 MST 구성을 위해서 다음의 원칙들이 기본적으로 적용된다.

첫째, 트리 구성에서 정점으로 표현되는 각 용어는 가장 가까이에 있는 용어 그룹에 연결된다.

둘째, MST의 부분트리(subtree)에 해당하는 고립된 용어그룹들은 가장 짧은 링크로서 결합될 수 있는 정점에 연결된다.

3.3. 결정함수

용어의존도에 의한 검색에서 최적의 검색규칙

으로 식(7)을 적용한다. 그리고 산출된 문헌검색가의 내림차순으로 검색결과를 제시하게 된다.

$$g(x) = \log \frac{p(x/rel)}{p(x/non\text{-}rel)} \quad \cdots\cdots\cdots\cdots (7)$$

결정함수 g(x)의 계산에서 각 문헌에 대한 적합성추정은 식(8)을 이용하게 된다.

$$p(w_1/x) = \frac{p(x/w_1)\ P(w_1)}{p(x)} \quad \cdots\cdots\cdots\cdots (8)$$

x는 문헌, w_1는 적합문헌을 의미한다.

4. 결 언

정보검색에서 용어의존도 모형이 보다 발전된 검색모형으로서 적용되기 위해서는 다음의 문제들이 해결되어야 한다.

첫째, 트리 모형에서는 직접연결된 용어쌍 이외에 간접적으로 연결된 용어 간의 관계에 의한 값은 거의 고려되지 않는다. 그러나, 실제적으로는 간접적인 용어 간의 관계가 존재한다.

둘째, BLE 모형에서는 계산을 효과적으로 하기 위해서 관련 용어쌍을 인위적으로 줄이면 검색효과에 영향을 주게 된다. 현재 수준에서는 질문에 관련한 용어 수(q)에 대해 q/2에 해당하는 용어쌍에 대한 계산을 하고 있다.

세째, 확장트리 모형에서는 연결된 충분한 수의 용어쌍에 의해 MST를 효과적으로 구축하는 방안이 모색되어야 한다.

네째, 질문이용에 의한 의존도 모형에서는 관련 용어들로 구성된 용어그룹 간의 의존도는 검색에서 고려되지 않았다.

참고문헌

정영미 & 노영희 (1992) "정보검색 기법의 검색효율 비교연구" 延世論叢 : 107-129.

Bookstein, A. (1983) "Information Retrieval: A Sequential Learning Process" JASIS. 34(5) :331-342.

Bookstein, A. (1983) "Outline of a General Probabilistic Retrieval Model" JD. 39(2) : 63-72.

Croft, W.B. (1986) "Boolean Queries and Term Dependencies in Probabilistic Retrieval Models" JASIS. 37(2) : 71-77.

Frakes, W.B. & Baeza-Yates, R.(1992) Infor- mation Retrieval : Data Structures& Algorithms. Englewood, Prentice-Hall.

Lam, K. (1982) "A Clustered Search Algorithm Incorporating Arbitrary Term Dependencies" ACM Transactions on Database Systems. 7(3) : 500-508.

Losee, R.M. (1988) "Parameter Estimation for Probabilistic Document-Retrieval Models" JASIS. 39(1) : 8-16.

Losee, R.M.(1994) "Term Dependence: Trun cating the Bahadur Lazarsfeld Expansion" IPM 30(2) : 293-303.

Maron, M.E. Kuhns, J.L.(1960) "On Relevance, Probabilistic Indexing and Information Retrieval" Journal of the ACM, 7(3), 216-243.

Robertson, S.E.(1990) "On Term Selection for Query Expansion" JD. 46(4) : 359-364.

Salton, G. (1989) Automatic Text Process ing: the Transformation, Analysis, and Retrieval of Information by Computer. Mass., Addison-Wesley.

van Rijsbergen, C.J.(1977) "A Theoretical Basis for the Use of Co-occurrence Data In formation Retrieval" JD. 33(2) : 106-119.

van Rijsbergen, C.J.(1979) Information Retrieval. London. Butterworths.

Yu,C.T., Buckley, C. Lam,K & Salton, G.(1983) "A Generalized Term Dependence Model in Information Retrieval" IT. 2 : 129-154.