3-계층 베이즈 네트워크를 이용한 정보검색 시스템 설계

○
서창석*, 이병욱*
*경원대학교 전자계산학과

Design of Information Retrieval System with 3-Tier Bayesian Networks

Chang-Seok Seo*, Byung-Wook Lee*
*Dept. of Computer Science, Kyungwon Univ.

요 약

수많은 주제와 수백만 개의 문서가 수록된 데이터베이스에서 사용자 질의에 적합한 문서를 찾는 과정에서 발생하는 불확실성 문제를 해결하기 위해 클러스터링 기법과 샘플링, 가능성과 확률 등 여러 가지 방법론이 연구되었다. 그 중 베이즈 이론에 근거한 베이즈 네트워크 모델이 제안되었다.

베이즈 네트워크는 원인과 결과를 연결한 네트워크로서 정보검색 시 발생하는 불확실성을 해결하는데 있어서 유용하며 다른 방법론에 비해 정보검색의 효율을 평가하는 기준인 조회율과 정확도에서 높은 효율을 보여준다. 그러나 베이즈 네트워크를 이용해 정보검색 시스템 구축 시 자식 노드가 많은 수의 부모 노드를 가질 경우에 확률계산을 위해 많은 링크 매트릭스가 필요하다는 단점이 있다. 본 논문에서 이러한 문제점을 해결하기 위해 2진 트리 분할 알고리즘을 사용하여 부모 노드를 특정 수의 그룹으로 분할하는 가상 네트워크를 포함하는 3-계층 베이즈 네트워크 모델을 제안한다.

1. 서론

정보검색 시스템은 1940년대 이후 발전되어 온 방대한 양의 도서를 관리하는데 도움을 주기 위해 개발되었다. 오늘날 많은 대학, 법인체, 공공 도서관은 단행본 도서, 정기간행물 등 여러 가지 문헌을 찾기 위해 정보검색 시스템을 사용하고 있다. 상용 정보검색 시스템은 수많은 주제에 있어서 수백만 개의 문헌이 수록된 데이터베이스를 제공한다. 이러한 환경 하에서 정보검색 시스템은 사용자가 다양한 문서들로부터 유용한 정보를 추출하기 위해 설계되었다.

정보검색에 관련된 업무는 문서와 질의 분석에 있어서 자연어의 모호성 때문에 불확실성을 포함하게 된다. 뿐만 아니라 실세계를 모델링 하는데 있어서 불확실성에 대한 문제는 계획, 추론, 문제해결, 의사결정 등 지능적인 행위가 필요한 모든 업무에서 존재한다.

이러한 불확실성을 해결하는데 있어서 널값(null values)을 이용한 방법론과 확실성 요소(certainty factors), 가능성과 확률(probability) [1] 등 현재까지 많은 연구들이 진행되어 왔다. 그 중 베이즈 이론(bayes' theory)에 근거한 베이즈 네트워크는 내용기반(content-based) 검색시스템 구현 시에 가장 유용한 모델로 사용되고 있다[2].

베이즈 네트워크(bayesian network)란 상호 배타적이고 완비적인 변수들을 연결한 그래프이다. 즉, 변수들의 집합 $\{x_1, x_2, x_3, \ldots, x_n\}$를 위한 베이즈 네트워크는 이러한 변수들의 결합확률분포(joint probability distribution)로 표현된다[3][4].

베이즈 네트워크는 두 개의 분리된 네트워크를 사용한다. 문서 집합에서 각각의 문서를 표현하는 문서 네트워크(document network)와 사용자의 질의를 표현하기 위한 질의 네트워크(query network)로 구

성된다. 문서 네트워크는 최소한 두 개의 노드 계층으로 구성되며 정적인 구조를 갖는 네트워크인 반면에 질의 네트워크는 사용자가 필요한 정보를 시스템에 알리는 역할을 수행하며 질의가 실행되는 동안에 동적으로 구축되는 네트워크이다.

베이즈 네트워크는 내용기반 정보검색에 있어서 불확실성을 처리하는데 유용한 해결책을 제공할 뿐만 아니라 기존의 다른 정보검색 시스템보다도 조회율(recall)과 정밀성(precision)에 관하여 높은 효율을 보여 준다[2]. 초기의 베이즈 네트워크의 주된 활용분야는 의학 진단 시스템이다. 그 후 베이즈 네트워크는 불확실성을 포함하는 많은 의사결정문제를 해결하는데 있어서 폭 넓게 사용되었다. 다른 추론 방법들과 비교해 볼 때 이러한 접근 방법은 의사결정 과정을 설명하는데 있어서 의미적인 풍부함과 불확실한 관계를 명확히 표현하는 장점이 있다.

그러나 베이즈 네트워크를 이용하여 정보검색 시스템을 구축하는 과정에서 자식 노드가 많은 수의 부모 노드를 가질 경우에는 사용자 질의에 적합한 문서를 매핑 하는 과정에서 계산이 복잡해지는 단점이 있다. 예를 들어, n개의 키워드를 갖는 문서를 검색하는데 있어서 2^n개의 링크 매트릭스가 필요하다[5]. 일반적인 문서에서의 키워드가 보통 20개 이상이라고 할 때, 사용자질의에 적합한 문서를 계산하는데 있어서 최소한 2^{20}개의 링크 매트릭스가 필요하다. 그 만큼 질의에 적합한 문서를 추론하는 과정에서 계산이 복잡해진다.

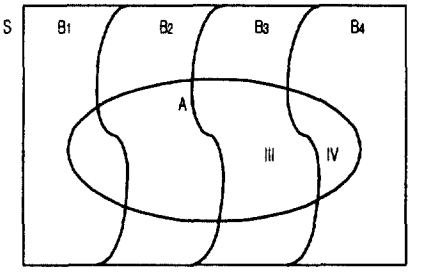
본 논문에서는 기존의 베이즈 네트워크 모델에 2진 트리 분할 알고리즘을 사용하여 부모 노드를 특정 수의 그룹으로 나누어 부모 노드와 자식 노드사이에 가상 네트워크를 추가한다. 정보검색 시스템 구축시 본 논문에서 제안한 2진 트리 분할 알고리즘을 적용하여 3-계층 베이즈 네트워크를 구축함으로서 매핑과정에서의 계산의 복잡성을 줄이고 검색성능을 향상시키는 모델을 제안한다.

2. 베이즈 네트워크

본 장에서는 정보검색을 위한 기본적인 모델과 확률이론에 근거한 베이즈 네트워크를 이용한 정보검색 시스템에 대해 살펴본다. 그리고 마지막으로 베이즈 네트워크의 문제점과 개선방향을 제시한다.

2.1 베이즈 정리(Bayesian' theorem)

베이즈 정리는 확률에 관한 계산과 이론을 기본으로 사전적 확률(prior probability)을 사후적 확률(posterior probability)로 수정하는데 사용된다. 수정은 보통 추가적인 정보가 있을 때만 가능한 것이며 새로운 정보나 표본들을 통하여 구할 수 있는 추가정보는 조건부확률을 구하는데 이용된다. 그림 2.1과 같이 표본공간이 B_1, B_2, B_3, B_4의 상호 배타적이고 완비적인(mutually exclusive and exhaustive) 사상들로 분할되어 있다고 할 때, 베이즈 정리는 다음과 같다[7].



* A의 면적 = I + II + III + IV

그림 2.1 베이즈 정리를 위한 벤다이어그램

$$P(B_2|A) = P(A \cap B_2) / P(A)$$
$$= P(A \cap B_2) / [\sum_{i=1}^{n} P(A|B_i)P(B_i)]$$
$$= P(A|B_2)P(B_2) / [\sum_{i=1}^{n} P(A|B_i)P(B_i)]$$

* B_2 대신 B_1, B_3, ···, B_n으로 대치가 가능하다.

2.2 베이즈 네트워크

베이즈 네트워크는 변수(사용자질의, 문서, 인덱스)들을 연결한 그래프이다. 이러한 변수들은 상호 배타적인 사상이다. 변수들의 집합 $\{x_1, , , \dots , x_n\}$의 베이즈 네트워크는 이러한 변수들의 결합확률분포로 표현된다. 불확실성을 처리하는데 있어, "원인" 변수들로부터 "결과" 변수들까지 호를 그려 베이즈 네트워크를 구축한다. 베이즈 네트워크에서 각각의 변수들은 노드로서 표현된다[6].

2.2.1 정의

집합 $\{x_1, \cdots, x_{i-1}\}$보다 작은 집합인 부모 집합 Π에 의해 조건적으로 설명될 수 있다. 이러한 집합이 주어졌을 때, 베이즈 네트워크는 비순환 방향그래프로 설명될 수 있다. 집합 $\{x_1, \cdots, x_n\}$에 대한 베이즈 네트워크는 다음과 같이 변수들에 대한 결합확률분포로 정의된다[6].

$$P(x_1, \cdots, x_n) = \Pi P(x_i|\Pi_i)$$

2.2.2 베이즈 네트워크 구조

베이즈 네트워크는 두 개의 분리된 네트워크를 사용한다. 문서의 집합에서 문서들을 표현하는 문서 네트워크(document network)와 사용자의 질의를 표현하기 위한 질의 네트워크(query network)로 구성된다[4].

문서 네트워크는 최소한 두 개의 노드 계층으로 구성된다. 최상위 계층은 키워드의 영역으로 구성된다. 최상위 계층은 텍스트 검색 시스템에서 문서집합에 알려진 모든 키워드를 포함하고 있기 때문에 이러한 계층은 문서집합의 사전을 표현한다. 다음 계층은 임의의 키워드의 조합으로 생성될 수 있는 개념, 사건, 또는 엔티티로 구성된다. 문서는 가능한 엔티티 중의 하나이다. 문서들과 키워드사이의 결합은 키워드에서 문서로의 직접 연결이 존재하는가에 따라 나타난다.

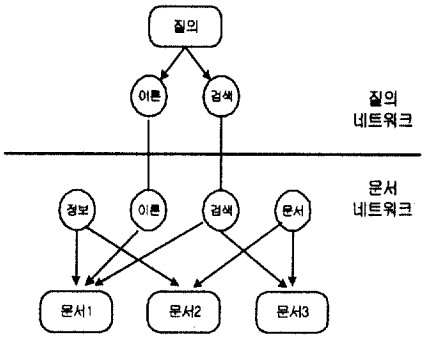


그림 2.2 베이즈 네트워크 모델

예를 들어, 그림 2.2에서 '문서1'은 키워드 '정보'와 '이론', '검색'으로 표현되고, '문서3'은 키워드 '검색'과 '문서'로 표현된다는 것을 쉽게 알 수 있다.

질의 네트워크는 사용자가 필요한 정보를 시스템에 알리는 역할을 수행하며 질의가 실행되는 동안에 동적으로 구축되는 네트워크이다. 그림 2.2에서는 키워드 '이론'과 '검색'으로 표현되는 사용자 질의를 보여 준다.

확률이론에 근거한 베이즈 네트워크의 장점은 불확실한 관계를 명확하게 표현하고 효율적인 추론 알고리즘을 제공하는 것이다.

2.3 베이즈 네트워크를 이용한 정보검색 시스템 설계의 문제점

베이즈 네트워크가 불확실한 관계를 명확하게 표현하고 효율적인 추론 알고리즘을 제공하는 장점이 있는 반면 하나의 자식 노드가 많은 수의 부모 노드를 가질 때, 확률계산과정이 복잡하다는 단점이 있다. 즉, 사용자 질의에 적합한 결과를 얻기 위해 많은 계산이 필요하다는 것을 의미한다.

하나의 자식 노드가 많은 수의 부모 노드를 가질 때 사용자 질의와 문서간에 매핑과정이 복잡하다. 예를 들어, 하나의 문서가 n개의 인덱스를 가지고 있을 때 질의에 적합한 문서를 찾기 위해서는 2^n개의 링크 매트릭스가 필요하다. 일반적인 경우에, 보통의 문서가 20개 이상의 인덱스를 가지고 있다면 2^{20}개의 링크 매트릭스가 필요하다. 그 만큼 매핑과정에 필요한 계산이 복잡해진다.

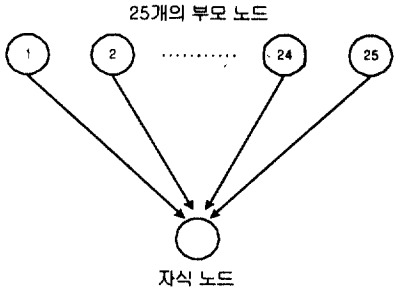


그림 2.3 베이즈 네트워크 모델의 문제점

그림 2.3은 25개의 부모노드를 갖는 자식노드에 대한 베이즈 네트워크 구조를 보여 준다. 이러한 구조에서는 링크 매트릭스가 2^{25}개가 필요하다. 그 만큼 네트워크 사이즈가 커지게 되고 확률계산과정이 복잡하다는 단점이 있다.

3. 3-계층 베이즈 네트워크 모델

본 장에서는 2장에서 살펴본 베이즈 네트워크 모델의 문제점을 해결할 수 있는 새로운 알고리즘을 설계한다. 베이즈 네트워크가 불확실한 관계를 명확하게 표현하고 효율적인 추론 알고리즘을 제공하는 장점을 활용하여 베이즈 네트워크 모델에 가상 노드를 포함하는 가상 네트워크를 추가함으로서 하나의 자식 노드가 많은 부모 노드를 가질 때, 발생하는 확률계산의 복잡성을 해결하는 3-계층 베이즈 네트워크 모델을 제안하고 가상 네트워크를 구축하는데 사용하는 2진 트리 분할 알고리즘을 제안하고자 한다.

3.1 3-계층 베이즈 네트워크

3-계층 베이즈 네트워크 모델의 구조는 기존의 베이즈 네트워크 모델의 구성요소인 질의 네트워크와 문서 네트워크와 본 논문에서 제안한 가상 네트워크(virtual network)로 구성된다.

가상 네트워크 구조는 하나의 자식 노드가 많은 수의 부모 노드를 가질 경우, 자식 노드와 부모 노드사이에 가상 노드를 생성하여 부모 노드를 m개 그룹으로 나누어 각 그룹과 가상 노드를 연결하고 가상 노드와 자식 노드를 연결하는 네트워크이다.

그림 3.1은 3-계층 베이즈 네트워크 모델의 예이다. n은 문서집합 U가 갖는 인덱스의 개수이고 m은 인덱스 수 n을 2진 트리 알고리즘을 사용하여 분할한 그룹의 수를 나타낸다.

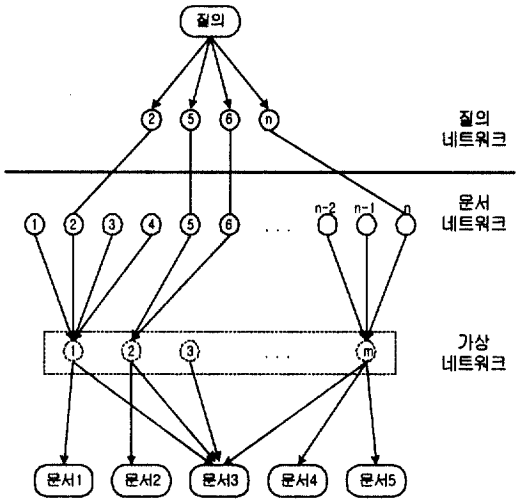


그림 3.1 3-계층 베이즈 네트워크 모델

3-계층 베이즈 네트워크를 이용한 정보검색 시스템의 장점은 사용자 질의에 적합한 문서를 찾는 과정에서 기존 베이즈 네트워크 모델보다 링크 매트릭스의 수가 감소한다는 장점이 있다. 링크 매트릭스의 감소는 곧 확률계산 과정의 빠르게 수행하게 됨으로 질의와 문서사이에 매핑시간을 감소시키는 결과를 갖는다.

3.2 2진 트리 분할 알고리즘

3-계층 베이즈 네트워크 모델에서 부모 노드를 분할하여 가상 네트워크를 구축하는 방법으로 2진 트리 분할 알고리즘을 제안한다. 제안한 알고리즘은 그룹의 개수를 결정하는 부분과 부모 노드가 갖는 전체 비트에서 비교할 비트의 개수를 결정하는 부분, 각각의 부모 노드를 각각의 그룹으로 할당하는 부분으로 나눌 수 있다. 본 절에서는 부모 노드를 그룹핑하는데 사용되는 2진 트리 분할 알고리즘의 구성과 알고리즘의 각 부분을 논한다.

2진 트리 분할 알고리즘
[N : 키워드(K) 개수, M : 그룹(G)의 개수, NG : 비교하는 키워드의 비트 수]

```
//그룹의 개수를 결정
for i = 1; to N
        if N <= 2^i then M = i; exit;
        endif
endfor

Min(NG) =< log2 M       // (M =<2^NG)

// 각각의 부모노드를 그룹에 할당
for n=1 to N
  for i=1 to NG
    키워드K(n)의 i번째 비트를 비교
  endfor [check for NG-bit data]

  switch (value of keyword)
    case 0 : apply K(n) to group G(0)
    case 1 : apply K(n) to group G(1)
                ...
    case M-1 : apply K(n) to group G(M-1)
    case M   : apply K(n) to group G(M-1)
                ...
  end switch
endfor
```

그림 3.2 2진 트리 분할 알고리즘

가상 네트워크를 추가하기 위한 첫 번째 단계로, 각각의 부모 노드를 분할하기 위한 그룹의 개수 M을 결정한다. 그룹의 개수는 N의 개수에 따라 변경된다. 그림 3.2는 그룹의 개수를 결정하고 각각의 부모 노드에 대해 비교할 비트 수를 계산하고 각각의 부모 노드의 시작 비트부터 비교할 비트까지의 값을 계산하여 각각의 그룹에 할당하는 알고리즘이다.

부모 노드를 그룹핑하는 그룹의 개수는 부모 노드의 개수에 따라 결정한다. 즉 하나의 자식 노드가 n개의 부모 노드를 갖는 경우, 그룹의 개수 M은 'n >= 2^m'을 만족하는 m에 의해 결정된다. 즉, 'm <= $\log_2 n$'의 최소 값이 그룹의 개수 M이다. 예를 들어, 그림 3.3은 25개의 부모 노드를 다섯 개의 그룹으로 분할한 예이다.

각각의 부모 노드를 그룹으로 할당하는 방법은 다음과 같다. 먼저 각각의 부모 노드에서 비교할 비트의 수를 결정한 다음, 부모 노드의 비트 값을 비교하여 그룹ID와 일치하는 그룹에 할당한다. 예를 들어, 그림 3.3에서 부모 노드 K_i = (0, 1, 0, ... , 1)의 값을 가질 때, K_i는 그룹 3에 할당된다.

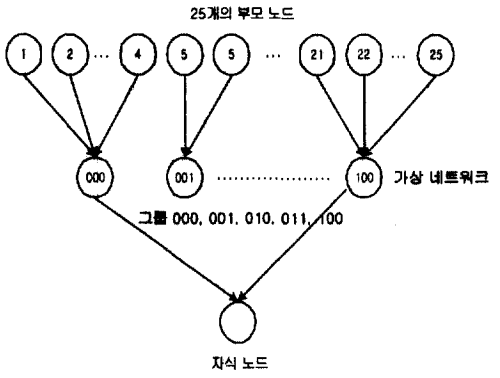


그림 3.3 3-계층 베이즈 네트워크의 그룹 개수

4. 성능평가

베이즈 네트워크를 이용한 정보검색 시스템 설계 시 성능을 평가하는 방법은 확률이론에 근거한 계산 방법과 모의 실험을 이용해 조회율과 정확도를 평가하는 방법으로 나뉘어 진다. 본 논문에서는 제안한 모델의 링크 매트릭스의 감소는 베이즈 네트워크의 분리기법을 이용해 증명하고, 기존의 베이즈 네트워크와 3-계층 베이즈 네트워크를 이용한 정보검색 시스템 구축 시 조회율과 정확도는 모의실험을 통해 비교·평가하고자 한다[9].

4.1 링크 매트릭스

그림 2.3의 베이즈 네트워크를 그림 3.3에 3-계층 베이즈 네트워크로 수정했을 때 링크 매트릭스 수의 감소를 베이즈 네트워크의 분리(divorcing)기법의 정의에 의해 링크 매트릭스의 감소를 증명할 수 있다. 그림 4.1의 모든 변수를 3진수로 가정했을 때, 그림 4.1의 a)를 표현하기 위해서 필요한 분포는 3^4개 즉, 81개가 필요한 반면에 분리기법을 이용한 b)에서의 분포는 $(3^2 + 3^3)$개 즉, 36개가 필요하다[6]. 즉, 모든 변수가 2진 값을 갖는 경우, n개의 부모 노드를 그림 3.3과 같이 m개의 그룹으로 분할했을 때 링크 매트릭스 수를 평균적으로 2^n개에서 $(\frac{n}{m}+1)\times 2^m$개로 줄일 수 있다.

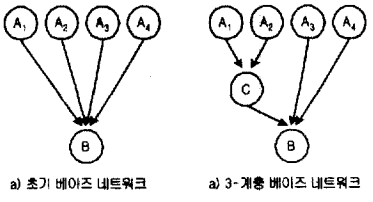


그림 4.1 베이즈 네트워크의 분리기법

4.2 모의실험 환경 및 모델

모의 실험 환경은 SUN UNIX에서 OS는 Solaris 1.1, 모의실험 도구는 SLAMⅡ을 사용하고자 한다.

두 모델의 조회율과 정확도를 평가하기 위한 모의 실험 모델의 구성은 그림 4.2와 같다. 키워드 생성기는 문서집합내의 문서를 표현하는 키워드를 생성하고 네트워크 생성기는 키워드를 이용하여 문서 네트워크를 구축한다. 질의 생성기는 사용자 질의를 이용하여 동적으로 질의 네트워크를 구축한다. 이러한 모델의 결과는 사용자 질의에 적합한 문서이다. 반환된 문서를 이용하여 조회율에 따른 정확도를 비교한다.

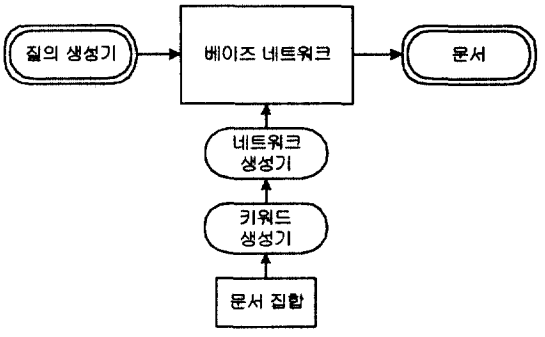


그림 4.2 모의실험 모델

모의실험에 사용되는 매개 변수는 문서, 각각의 문서가 포함하는 인덱스 개수와 질의이다. 본 논문에서는 인덱스의 개수가 5에서 20개인 100개의 문서와 인덱스의 개수가 20에서 50개인 1000개의 문서를 대상으로 모의실험을 수행하며, 질의는 주어진 인덱스에 대해 임의적으로 생성하여 각 질의에 대한 적합한 문서를 추출하여 조회율에 따른 문서의 정확도를 평가했다.

4.3 모의실험 결과

모의실험 결과는 사용자 질의에 결과인 문서를 대상으로 베이즈 네트워크와 3-계층 베이즈 네트워크 모델을 평가한다.

표 4.1은 문서집합의 개수가 100개일 경우에 조회율에 따른 정확도의 변화를 비율을 나타낸다. 이 경우에 있어서 평균 4%정도의 정확도의 향상을 보이나 문서집합의 크기가 1000인 경우에 있어서는 평균 9% 정도의 정확도의 향상을 보인다.

그림 4.3은 문서집합의 개수가 1000이고 각각의 문서가 가지는 인덱스의 개수가 20에서 50일 경우의 조회율에 따른 정확도를 나타낸다.

표 4.1 문서집합의 개수가 100개일 경우

조회율	정확도(%)	
(%)	BN	3-Tier BN
10	68.12	72.59
20	67.72	69.56
30	61.30	65.47
40	57.53	59.60
50	53.45	54.85
60	48.85	51.70
70	35.98	42.40
80	28.08	37.56
90	25.90	31.59
100	23.85	27.58
평균	47.08	51.29

표 4.2 문서집합의 개수가 1000개일 경우

조회율	정확도(%)	
(%)	BN	3-Tier BN
10	63.22	69.33
20	61.55	67.76
30	52.23	65.82
40	48.35	59.68
50	46.24	55.95
60	41.75	53.71
70	34.88	41.39
80	26.18	36.49
90	23.90	33.55
100	21.85	28.98
평균	42.02	51.27

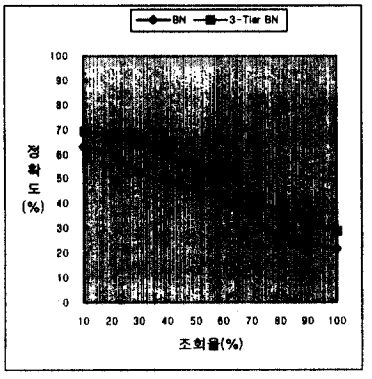


그림 4.3 BN와 3-Tier BN의 조회율과 정확도 관계

5. 결 론

텍스트-기반 데이터베이스 시스템에서 질의를 처리하는데 있어서 불확실성 문제를 해결하는데 있어서 많은 연구가 진행되어 왔다. 이러한 연구들 중에 확률이론에 근거한 베이즈 네트워크 모델은 다른 연구방법론들의 비해 추론과정을 명확히 설명하고, 조회율과 정확도에 있어서도 높은 성능을 보여 준다.

그러나 베이즈 네트워크를 구축 시 자식 노드가 많은 수의 부모 노드를 가질 경우에 사용자 질의에 적합한 문서를 매핑하는 과정이 매우 복잡함으로 이런 경우에 베이즈 네트워크를 정보검색 시스템에 적용하는 것은 적합하지 않다.

본 논문에서 제안한 3-계층 베이즈 네트워크는 문서집합의 갱신이 자주 일어나는 경우에 가상 네트워크를 재 구축해야 하는 단점이 있으나, n개의 부모 노드를 m개의 그룹으로 분할했을 때 링크 매트릭스 수를 평균적으로 2^n개에서 $(\frac{n}{m}+1)\times 2^m$개로 줄일 수 있다. 즉, 하나의 자식노드가 많은 수의 부모 노드를 가질 때 발생하는 확률계산의 복잡성을 상당히 줄일 수 있다. 또한 정보검색의 효율을 측정하는 조회율과 정확도에 있어서도 7% 이상의 성능 향상을 보였다.

참고문헌

[1] K.L Kwok,"Network Approach to Probabilistic Information Retrieval," ACM Transaction on Information Systems, Vol 13, pp 324-353, 1995.

[2] M.T.Indrawan, B.Srinivasan "Optimising Bayesian Belief Networks: A Case Study of Information Retrieval Systems," Proceedings of the 1998 IEEE-Volume 3 , pp2273-2278, 1998.

[3] David Heckerman, Michael P. Wellman "Bayesian Networks," ACM Vol.38. No. 3, 1995.3.

[4] J. Pearl "Bayesian Networks," MIT Press, pp 149-153, 1995.

[5] Judea Pearl "PROBABILISTIC REASONING IN INTELLIGENT SYSTEMS" Morgan Kaufmann , pp184-187, 1988.

[6] Finn V. Jensen "An Introduction to Bayesian Networks", Springer, pp7-64, 1996.

[7] 유지성 "현대통계학" 전영사, pp 52-70, 1998.

[8] Sheldon Ross, 우정수 역 "A First Course in Probability" 자유아카데 미, pp 71-112, 1994.

[9] 류근호, 김진호 공역 "정보검색", 시그마프레스, pp 509-567, 1995.