퍼지 필터링 구조를 이용한 멀티미디어 통계 사서함 시스템

이 종 득† · 김 대 경††

요 약

최근에 인터넷과 통신망의 활성화로 인하여 특정 도메인 정보들을 효율적으로 구축하고 서비스하기 위한 여러 가지 방법들이 제안되고 있다. 본 논문에서는 멀티미디어 통계 사서함 구축을 위한 퍼지 필터링 객체 관련성 기법을 제안한다. 제안된 기법은 ∂-cut을 이용하여 문서 객체들을 그룹화하기 위해 $RelCRO(D_{omain}, G_i)$, $RelSRO(D_{omain}, G_i)$, FAS를 이용한다. 또한 제안된 기법의 성능을 알아보기 위해 1600개의 멀티미디어 타입정보를 이용하여 실험을 수행하고 랜덤 키, OGM, 그리고 제안된 방법을 비교 분석한다. 그 결과 제안된 방법의 성능이 보다 효율적임을 보인다.

Multimedia Statistic Post-office Box System using Fuzzy Filtering Structures

Chong Deuk Lee† · Dae Kyung Kim††

ABSTRACT

According to the current increase of the usefulness of information by Internet Communication network, several methods are proposed in which a specific domain information may be efficiently constructed and serviced. This paper proposes Relationship Grouping of Fuzzy Filtering Objects for Multimedia Statistic Post-office Box Construction. The proposed method exploits $RelCRO(D_{omain}, G_i)$, $RelSRO(D_{omain}, G_i)$ and FAS in order to group using ∂-cut. To know how well the proposed method is able to work, this paper have test against the methods with 1600 items of multimedia type information, and our system are compared with Random-Key, OGM and the proposed method. The results shows that the proposed method provides the better performance than the other methods.

키워드 : 퍼지 필터링(Fuzzy Filtering), ∂-cut, 통계 사서함(Statistic Post-Office Box), 그룹화(Grouping)

1. 서 론

최근에 정보이용의 확대로 인하여 멀티미디어 통계데이터 서비스는 웹 상에서 중요한 정보로 떠오르고 있다.

통계 문서정보는 멀티미디어 데이터 형태로 관리되고 서비스됨에 따라 텍스트, 오디오, 비디오 정보와 같은 멀티미디어 정보들을 관리하기 위한 새로운 기법이 제안되고 있다[2-5, 13].

지금까지 통계 문서 정보 서비스는 물가 통계, 환경통계, 지역통계, 설비투자, 추계지수, 서비스업 활동 지수, 장래가구 추계 작성 등의 이용자 요구에 부응하는 텍스트 중심의 서비스가 주를 이루고 있으며, 도메인 중심의 문서 정보를 그룹화하여 서비스하는 멀티미디어 문서 정보 서비스는 미흡한 실정이다[6, 7, 10, 11].

일반적으로 응용 도메인의 문서객체를 분류기법에 따라 그룹화 하기 위해서는 응용도메인의 문서 객체를 탐색해야 되며, 탐색 결과에 따라 그룹화 여부가 결정되게 된다. 이러한 기법을 위해 [11]은 OGM(Optimistic Genealogy Method), BGM(Balanced Genealogy Method), RGM(Recursive Genealogy Method) 기법들을 이용하여 카테고리를 단순하게 생성하는 기법을 제안하였으며, [8]에서는 퍼지 관계와 통계적 클러스터링 기법을 이용한 문서 구조화 방법을 제안하였다. 그러나 이러한 기법들은 비용이 많이 들며, 응용 도메인의 문서 정보가 대규모로 증가할 때는 효율적이지 못하다는 문제점이 제기되고 있다.

현재 사용자 중심의 통계정보 서비스는 많이 개선되고 있지만 텍스트 중심의 문서 서비스에서 벗어난 멀티미디어 문서 정보서비스, 체계적인 관리, 모집단의 관리, 조사의 관리 및 서비스, 서비스 정보 분류는 상당히 미흡한 실정이다. 이러한 이유는 지금까지 통계정보는 단순 텍스트 정보로만 인식되어 왔기 때문이며, 이용되는 통계 정보 또한 단순 데이터 및 정보로만 인식되어왔기 때문이다. 이러한 문

※ 본 논문은 정보통신부 연구 진흥원에서 지원하고 있는 정보통신 기초기술 연구지원 사업의 연구결과입니다.
† 정 회 원 : 국립 익산대학 정보통신공학과 교수
†† 정 회 원 : 전북대학교 통계정보학부 교수
논문접수 : 2004년 2월 19일, 심사완료 : 2004년 7월 14일

제를 해결하기 위해서 통계청에서는 STAT-KOREA 시스템 등을 구축하여 정보 서비스를 수행하고 있으나 통계 정보들을 단순하게 나열하는 것만으로는 통계 문서 서비스 개선이 이루어지지 않으며, 다양한 데이터 형태의 산재되어 있는 통계 정보들을 서로 의미적으로 결합하여 구축하기란 쉬운 일이 아니다[1, 3, 4].

따라서 본 논문에서는 통계정보의 양이 커지고 사용자가 요구하는 정보를 선별적으로 서비스할 수 있도록 퍼지 필터링 객체 관련성 구조를 이용한 사서함 시스템을 제안하며, 제안된 방법은 단순 구조화 방법과는 다르게 통계정보들을 개념적, 의미적 객체 단위로 구성하여 구축하는 방법으로서 응용도메인에서 제공된 객체와 객체 프로파일 정보를 이용하여 시스템을 구축하게 된다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련연구를 알아보며, 3장에서는 멀티미디어 통계 사서함 구축 모델을 제안하며, 4장에서는 제안된 기법에 대해서 실험적 평가를 수행하며, 끝으로 결론 및 향후 연구방향에 대해서 알아본다.

2. 관련 연구

이 절에서는 멀티미디어 통계 사서함을 구축하기 위해 사용되는 기법인 통계적 클러스터링 기법, Genealogy Measure 기법 등에 대해서 살펴보기로 한다.

2.1 통계적 클러스터링 기법

통계적 클러스터링 기법[8, 12]은 분할적 클러스터링 기법과 계층적 클러스터링 기법으로 구분되며, 분할적 클러스터링 기법은 중첩된 분할 구조가 아닌 평평한 하나의 분할 구조로 클러스터링을 구성해 나가는 기법이다. 그리고 계층적 클러스터링 기법은 가장 유사한 두 개체를 선택하여 병합해 나가는 기법으로서 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법 등으로 구성된다.

계층적 클러스터링은 각각의 모든 문서 데이터 개체가 최하위 계층구조에서 최상위 계층구조로 클러스터링이 구성되게 하는 방법이며, 분할적 클러스터링은 유클리디안 거리(Euclidean Distance)를 이용하여 가까운 문서들을 클러스터링으로 구성하는 방법으로서 차원의 제약이 없고 간단하다는 장점으로 널리 사용되고 있다. 그러나 이 기법은 같은 타입을 가진 인스턴스 객체들에 대해서는 적합한 구조화 방법이지만 타입들이 분산 형태로 구조화되어 있을 경우에는 객체 관계성 파악이 어렵다는 문제점이 발생되고 있다.

2.2 Genealogy Measure 기법

Genealogy Measure기법[11]은 각 문서 객체에 대한 유사도를 계산하는 기법으로서 OGM, BGM, RGM 기법으로 구분된다.

2.2.1 OGM 기법

이 기법은 하나의 그룹의 각 원소들에 대해서 유사도를 계산하는 방법이며, 그리고 난 후 각 원소들에 대한 평균 가중치를 구하여 그룹들간에 대한 유사도를 계산한다. 이때 원소의 유사도는 해당 원소가 다른 그룹과 얼마나 잘 일치(matching)되느냐를 결정하게 된다. 그룹에서의 원소에 대한 유사도 계산은 $leafsim_{(G1, G2)}(k_i) = depth(LCA_{(G1, G2)}(k_i))/depth(k_i)$이다.

여기서 $leafsim_{(G1, G2)}(k_i)$는 노드 k_i가 그룹 G1, G2에 있는 원소들과 얼마나 잘 결합되는지를 나타내는 유사도이며, LCA (Lowest Common Ancestor)는 최대 깊이를 가진 노드이다. 유사도는 [0, 1] 사이이며, G1과 G2안에 최대 유사도 유지하면 $leafsim_{(G1, G2)}(k_i) = 1$이고, 그렇지 않으면 $leafsim_{(G1, G2)}(k_i) = 0$이다. OGM 기법은 그룹의 원소들에 따라 유사도가 결정되게 된다. 이 기법은 유사도를 이용한 문서 구조화 방법으로 최적의 기법을 제공하지만 평균 가중치를 구하여 문서의 일치성(matching)을 검사하여 결합성을 측정해야 하는 문제점이 발생되고 있다.

2.2.2 BGM 기법

이 기법은 OGM에서 발생되는 유사도 중복 문제를 해결하기 위한 기법으로서 [0, 1] 사이의 매개변수(parameter) β를 이용하여 결합성의 다양성(multiplicity)으로 인한 유사도가 떨어지는 문제를 해결하기 위한 기법이다. 그러나 이 기법은 [0, 1] 사이의 매개변수를 이용하여 결합성의 문제를 부분적으로 해결하였지만 같은 그룹에서 같은 타입을 가지고 있는 문서들에 대해서 결합성을 반복하여 측정해야 하는 문제점이 발생되고 있다.

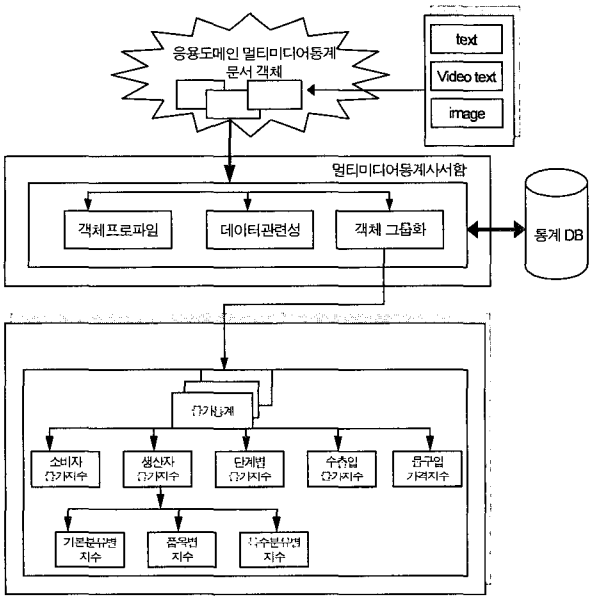
2.2.3 RGM 기법

이 기법은 다중 발생에 관한 문제를 해결하기 위한 기법으로서 하나의 트리 구조에서 많은 노드로 구성되어 있을 때 유사도의 중복성 수행으로 인한 문제점을 개선하기 위한 기법이다. RGM 기법은 트리 구조에 대한 가중치를 이용하여 그룹을 구성하는 기법으로서 원소 n에 대하여 임의의 그룹 G의 잎 노드와 내부노드로 구분하여 그룹화를 구성하게 되며 잎 노드일 때는 WG(n)=W(n), 내부 노드일 때는 $\min(\sum_{n \in G} W_G(n))$으로 표현된다. 노드 n이 G의 원소가 아닐 때는 가중치는 0으로 표현되며, 그룹의 원소에 대한 가중치는 θ로 표현된다.

이와 같은 방법은 원소의 각 유사도에 따라 문서들이 그룹화 될 수 있도록 하는 방법이며, 관련 문서들이 유사도에 따라 퍼지 필터링이 수행될 수 있도록 하기 위한 방법이다. 그러나 이 기법은 그룹화 구조에서 많은 문서 노드가 구성되어 있을 때는 유사도의 중복 수행이 발생되는 문제점을 가지고 있다.

3. 멀티미디어 통계 사서함 구축 모델

본 논문에서 제안된 통계 사서함 구조는 시스템과 이용자들이 응용도메인에서 제공된 문서정보를 서비스 받을 수 있도록 구성하며, 제안된 구조는 응용 도메인으로부터 추출된 객체를 분류하기 위한 객체 프로파일 구조와 추출된 객체 프로파일을 이용하여 이들의 관계성을 개념적 관계와 의미적 관계를 파악하기 위한 데이터 관계성으로 구성된다. 그리고 이들과 상호작용을 통해서 객체를 분류하여 객체 클래스 단계를 구성하는 객체 그룹화구조로 구성되며 제안된 시스템 구성도는 (그림 1)과 같다.



(그림 1) 통계 사서함 시스템 구성도

3.1 객체 프로파일

응용 도메인 공간 $S=(D_{omain}, p)$는 객체들의 응용도메인 D_{omain}와 프로파일 p로 정의되며, 응용 도메인 공간상에서 객체 $(O_x, O_y, O_z \in D_{omain})$의 조건은 다음과 같은 성질을 만족한다.

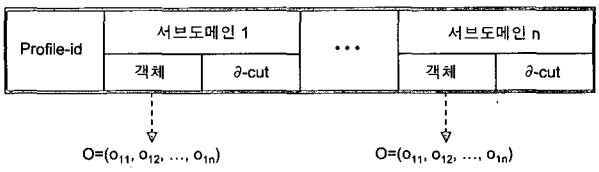
(1) $p(O_x, O_y) = p(O_y, O_x)$
(2) $0 \le p(O_x, O_y) \le 1,\ O_x \neq O_y$
(3) $p(O_x, O_x) = 1$
(4) $p(O_x, O_y) \le p(O_x, O_z) + p(O_z, O_y)$

위와 같은 성질을 만족하고, 응용 도메인 공간 S가 주어질 때 응용 도메인 공간상에서 서브 도메인 SD_{omain}은 다음과 같이 정의된다.

[정의 1] $SD_{omain} = SD_{omain}(O_x,\ r_x) = \{O_i \in D \mid p(O_x,\ O_i) \le r_x\}$이다.

여기서 r_x는 서브 도메인에서 중심객체 O_x를 중심으로 한 반경 안의 거리이며, SD_{omain}은 O_x와 반지름 $r_x \ge 0$ 의해 결정된다.

이와 같은 조건을 만족할 때 객체 관계성을 결정하기 위한 객체 프로파일의 구조는 (그림 2)와 같다.



(그림 2) 객체 프로파일 구조

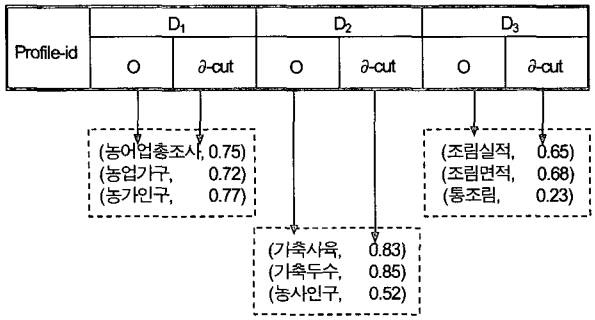
(그림 2)에서 객체 $(O = o_{11}, o_{12}, \cdots, o_{1n})$은 해당 서브 도메인에서 추출된 프로파일 객체로 구성되며, ∂-cut은 객체 관련성을 나타내는 퍼지 필터링이다[8].

퍼지 필터링은 객체 관련성을 [0, 1] 사이의 값에서 임의의 $\partial (0 \le \partial \le 1)$ 값이 소속되는 정도를 퍼지 집합으로 표현한 값으로서 다음과 같이 정의된다.

[정의 2] $\partial\text{-}cut = \{O \mid D_{omain}(o) \ge \partial\}$이다.

여기서 $D_{omain}(o)$는 도메인 D_{omain}에서의 이의의 객체정보이다.

예를 들어 3개의 멀티미디어 응용도메인 D_{omain1}, D_{omain2}, D_{omain3}에서 D_{omain1}의 객체 프로파일 정보는(농업총조사, 0.75), (농업가구, 0.72), (농가인구, 0.77), D_{omain2}에서 객체 프로파일 정보는 (가축사육, 0.83), (가축두수, 0.85), (농사인구, 0.52), D_{omain3}에서 객체 프로파일 정보는(조림실적, 0.65), (조림면적, 0.68), (통조림, 0.23)이라고 하면, (그림 2)를 이용한 객체 프로파일 구조는 (그림 3)과 같다.



(그림 3) (그림 2)를 이용한 객체 프로파일 구조

(그림 3)의 ∂-cut은 객관적인 프로파일 객체를 기술해 놓은 퍼지 관련성으로서 D_{omain3}=(통조림, 0.23)의 경우 D_{omain3}에 있음에도 불구하고, 퍼지 관련성이 낮음을 알 수 있다. 퍼지 관련성은 0에서 시작하며, 갱신 과정을 거쳐 0과 1사이의 값으로 표현된다.

3.2 객체 관련성

이 절에서는 객체 관련성에 따라서 문서 객체들을 그룹화

하기 위한 *CRO*와 *SRO*에 대한 관련성과 *JM*(Joint Matrix) [5, 11] 관련성, 그리고 *FAS*(Fuzzy Average Similarity)에 대해서 살펴본다.

3.2.1 *CRO*와 *SRO* 관련성

문서 객체 관련성을 위한 응용도메인의 문서 도메인을 D_{omain}라하고, n개의 그룹을 G_1, G_2, …, G_n이라 하자. 이때 도메인 D_{omain}에서 그룹화를 수행할 객체관련성은 *CRO* 분류기법과 *SRO* 분류기법으로 구분하여 분류한다.

첫째 *CRO*(*Conceptual Reference Object*) *분류기법* - 이 기법은 응용 도메인 D_{omain}가 분류할 충분한 문서를 가지고 있을 때 문서 객체를 분류하는 기법으로서 모든 문서 객체 그룹에 문서를 할당하는 기법이다.

둘째 *SRO*(*Semantic Reference Object*) *분류기법* - 이 기법은 의미적으로 연관관계를 가진 문서객체들을 분류하는 기법으로서 모든 문서 객체 그룹에 의미적 관련성을 가진 문서객체들을 할당하는 기법이다.

*CRO*는 보다 확장된 영역에서의 다양한 문서 객체를 개념적으로 분류하기 위한 문서 객체 분류 기법이며, 응용도메인의 문서객체 도메인을 D_{omain}라하고, 문서객체 분류를 위한 그룹을 $G_i \in G$라 할 때, 도메인에서의 *CRO*의 분류관계는 $CRO_{number}(D_{omain}, G_i)$이며, 이때 G_i는 개념적으로 관련된 문서 객체의 수이다. 그리고 *SRO*는 의미적으로 관련이 있는 문서 객체 분류이며, 도메인에서의 *SRO* 분류 관계는 $SRO_{number}(D_{omain}, G_i)$이며, G_i는 의미적으로 관련된 문서 객체의 수이다.

이때 *CRO*와 *SRO* 관계에 있어서 *SRO*에 대한 도메인 관계는 다음과 같이 정의된다.

[정의 3] $SRO_{number}(D_{omain}, C)=CRO_{number}(D_{omain}, C_i)/|D_{omain}|$ *이다.*

여기서 $|D_{omain}|$는 응용 도메인에서의 문서 객체의 수(size)이다.

*SRO*의 임계값은 0과 1사이의 퍼지 관련성이며, 분류기법에 따라 문서 객체들이 그룹화되게 된다. 본 논문에서 임계값을 위해 [정의 2]에서 정의된 퍼지 필터링 ∂-cut을 이용하며, *SRO*에 대한 퍼지 필터링과 *CRO*에 대한 퍼지 필터링은 다음과 같이 정의된다.

[정의 4] $\partial\text{-}cut_{(CRO)} = \{O \mid D_{omain}(o) \geq \partial\}$이다.

[정의 5] $\partial\text{-}cut_{(SRO)} = \{O \mid D_{omain}(o) \geq \partial\}$이다.

따라서 퍼지 필터링을 만족하는 문서 객체들이 분류 기법에 따라 관련성을 갖기 위해서는 다음과 같은 조건을 만족해야 한다.

[조건 1] $SRO(D_{omain}, G_i) \geq \partial\text{-}cut_{(SRO)}$

[조건 2] *G_i의 모든 조상(supergroup) G_j에 대해서* $SRO(D_{omain}, G_j) \geq \partial\text{-}cut_{(SRO)}$

[조건 3] *G_i의 자식(subgroup) G_k에 대해서* $SRO(D_{omain}, G_i) \leq \partial\text{-}cut_{(SRO)} \leq 1$

이러한 조건을 이용하여 본 논문에서는 관련된 문서 객체를 그룹화하기 위하여 ∂-*cut*의 퍼지 필터링을 0.5 이상으로 설정하며, 퍼지 필터링을 만족하는 개념적 참조 문서 객체의 관련성 $RelCRO(D_{omain}, G_i)$*와* 의미적 참조 문서 객체 관련성 $RelSRO(D_{omain}, G_i)$는 다음과 같이 정의된다.

[정의 6] $RelCRO(D_{omain}, G_i) = \sum_{G=Gi} Rel(DO)$이다

여기서 *REL*은 *Relevance*이며, $RelCRO(D_{omain}, G_i)$는 그룹화 조건을 만족하는 개념적 참조 문서 객체 관련성을 가진 모든 문서의 수이다.

[정의 7] $RelSRO(D_{omain}, G) = RelSRO(D_{omain}, parent(G_i)).\ RelCRO(D_{omain}, G_i) / \sum_{G\text{는부모}(G_i)\text{의자식}} (RelCRO(D_{omain}, G_i))$이다.

의미적 참조 문서 객체 관련성은 개념적 참조 문서 객체의 관련성에 비해 좀 더 구체적인 관련성을 지닌 문서를 그룹화하기 위한 관계성이며, 의미적 참조 문서 객체의 관련성 *RelSRO* 표현은 다음과 같이 정의된다.

[정의 8] $RelSRO(D_{omain},) = \{RelSRO(D_{omain}, G_1), RelSRO(D_{omain}, G_2), \cdots, RelSRO(D_{omain}, G_n)\}$

예를 들어 4개의 응용 도메인 그룹 G_1 = "농림어업통계", G_2 = "물가통계", G_3 = "환경통계", G_4 = "전자상거래"에 대해서 질의 "농업 OR 물가 OR 환경 OR 전자상거래"를 수행한 결과 관련 문서 객체가 566이고, "농업 OR 물가 OR 소비자물가 OR 생산자 물가"를 수행한 결과 관련 문서 객체가 324, "생산비 OR 구매자 OR 경기지표"를 수행한 결과 관련 문서 객체가 65개이라면 의미적 참조 문서 객체의 수는 $RelSRO(D_{omain})$= (566, 324, 65)가 되며, 각각의 카테고리에 대한 의미적 참조 문서 객체의 관련성 $RelSRO(D_{omain})$은 $RelSRO(D_{omain})$ = {566/(566+324+65), 324/(566+324+65), 65/(566+324+65)} $RelSRO(D_{omain})$ = { 0.60, 0.34, 0.07 }이 된다.

3.2.2 JM 관련성

*JM*관련성은 응용 도메인 D_{omain}에서 문서 객체 분류를 위한 문서 분류기에서 문서들을 잘 못 분류하는 문서 오류를 최소화하는 기법으로서 본 논문에서는 분류오류를 최소화하기 위해 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$를 이용한 결합 관련성 행렬 *JM*[12]을 제안하며, *JM*은 다음과 같이 정의된다.

[정의 9] 그룹 G_1, G_2, …, G_n에 대한 *NJM*(*Normalized Joint Matrix*) $JM = (m_{ij})$는 $n \times n$ *matrix*이다. 여기서 *matrix* m_{ij}

는 그룹G에서 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$를 수행 한 후 관련된 문서객체의 총합이다.

이때 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$를 결합한 결합 행렬은 다음과 같이 정의된다.

[정의 10] $(m_{ij}) \times (D_{omain}, G_i) = RelCRO(D_{omain}, G_i)$이다.

예를 들어 3개의 응용 도메인 그룹 G_1 = "농림어업통계", G_2 = "물가통계", G_3 = "환경통계"에서 2,000개의 $CRO(D_{omain})$의 문서 객체가 존재한다고 할 때 그룹 G_1과 관련된 CRO는 1000개, G_2와 관련된 CRO는 700개, G_3와 관련된 CRO는 300개라 하자. 이때 G_1과 관련된 SRO는 각각 600, 100, 50이고, G_2와 관련된 SRO는 400, 200, 70, G_3와 관련된 SRO는 250, 200, 100이라고 하면 JM은

$$JM = \begin{bmatrix} 600/1000 & 400/700 & 250/300 \\ 100/1000 & 200/700 & 200/300 \\ 50/1000 & 70/700 & 100/300 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.57 & 0.83 \\ 0.1 & 0.29 & 0.67 \\ 0.05 & 0.1 & 0.33 \end{bmatrix}$$ 이 되며, 따라서

$$JM \times (D_{omain}, G_i) = \begin{bmatrix} 600/1000 & 400/700 & 250/300 \\ 100/1000 & 200/700 & 200/300 \\ 50/1000 & 70/700 & 100/300 \end{bmatrix} \times \begin{bmatrix} 1000 \\ 700 \\ 300 \end{bmatrix} = \begin{bmatrix} 1250 \\ 500 \\ 220 \end{bmatrix}$$ 이 된다.

3.2.3 FAS

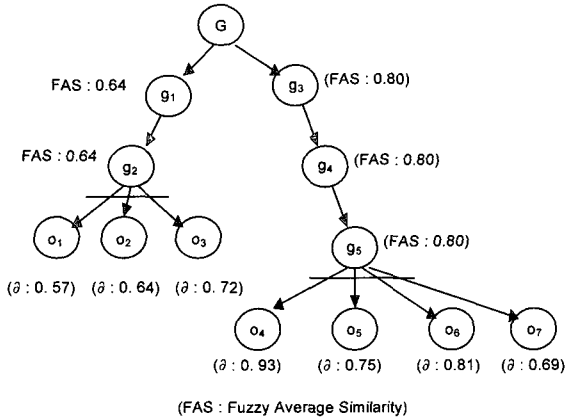
평균 퍼지 관련성 유사도 FAS는 문서 객체 그룹의 각 문서 객체들에 대해서 유사도를 계산하는 방법이며, 각 문서 객체원소들에 대한 평균 퍼지 관련성을 구하여 두 그룹 구조에 대한 유사도를 계산한다. 이 기법은 해당 문서 객체가 다른 문서 객체 그룹의 집합구조와 얼마나 잘 일치(matching)되느냐를 결정하게 되며, Genealogy Method[11]의 BGM, RGM에서 발생되는 유사도 중복성 문제를 해결하기 위해 사용되는 기법이다.

임의의 문서 객체 그룹 G와 G 구조 안의 임의의 문서 객체 O에 대해서 $C_G(n)$은 G의 모든 자식 노드들의 집합이라 하고, $W_G(O)$는 임의의 그룹 G의 문서 객체 O에 대한 가중치라 하자.

이때 문서 객체 그룹 G 구조에 대한 평균 퍼지 관련성 유사도 FAS는 다음과 같이 정의된다.

[정의 11] $FAS = \sum_{G_j\text{는부모}(G_i)\text{의자식}} (\partial\text{-}cut_{(SRO)}=\{O \mid D_{omain}(o) \geq \partial\}) / W_G(O)$이다.

예를 들어 (그림 4)에서 g_2에 대한 $D_{omain}(o) \geq 0.5$를 만족하는 문서 객체가 o_1, o_2, o_3이고 g_5에 대한 $D_{omain}(o) \geq 0.6$을 만족하는 문서 객체가 o_4, o_5, o_6, o_7이라면, g_2에 대한 $W_G(O)$는 3이며, g_5에 대한 $W_G(O)$는 4이다.



(그림 4) FAS 구조

따라서 [정의 11]에 의해 g_2에 대한 FAS는 0.64이며, g_5에 대한 FAS는 0.80이 된다.

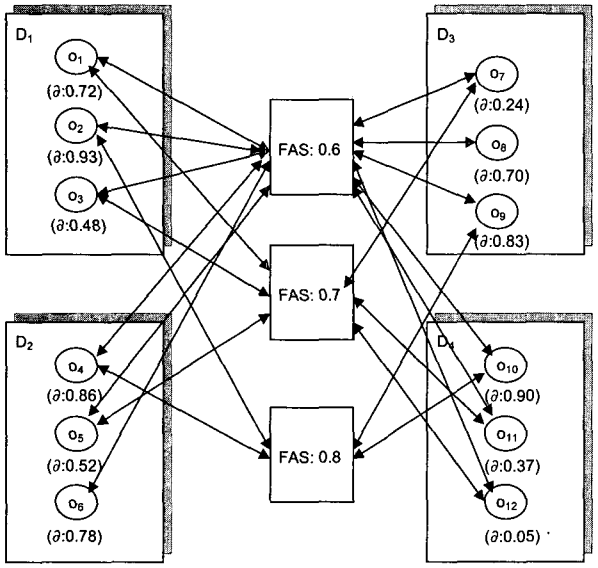
3.3 문서 그룹화

문서 객체 그룹화 생성과정은 데이터 관련성에 의해서 레퍼런스 된 문서 객체들의 퍼지 관련성을 수행하여 객체 그룹화를 생성하게 된다. 이러한 방법은 서로 관련 있는 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$의 문서 객체정보들을 효율적으로 구성하고 관리하기 위한 방법이며, 각 문서 객체 정보가 가지고 있는 프로파일정보를 이용하여 그룹화를 생성하게 된다.

이처럼 응용 도메인에서 FAS에 따라 $RelSRO(D_{omain}, G_i)$와 관련된 객체 그룹화를 생성하고, 객체 그룹화로 구성된 각 객체 정보들 간의 FAS와 $SRO(D_{omain}, G_j) \geq \partial\text{-}cut_{(SRO)}$인 객체 정보들에 대해서 그룹화를 생성하는 과정은 다음과 같이 진행된다.

```
단계 1 문서 객체정보들간의 FAS 수행
procedure fuzzy_value relation()
{
 procedure fuzzy_relation(oi, oj)
// 객체정보들 간의 퍼지 관련성을 수행한다.
procedure extend_search(ok)
// 객체 정보들간의 탐색을 확장한다.
   for (i = 0 ; i++)
     fuzzy filtering(oi, oi+1)
     extract ∂_value()
}
단계 2 FAS를 만족하는 객체 그룹 생성
procedure extend_search()
{
extract_∂(membership function)()
// 임의의 문서 객체 정보들에 대해서 퍼지 관련성 값을 생성한다.
for (i = 0 ; i++){
   if ((SRO(Domain, Gj) ≥ ∂-cut(SRO))
   create FAS Group(oi)
// 임의의 문서 객체정보들에 대해서 FAS를 만족하는 그룹을 생성한다.
}
```

예를 들어 (그림 5)와 같이 응용 도메인 D_1, D_2, D_3, D_4에 12개의 문서 객체타입 (o_1, ∂ : 0.72), (o_2, ∂ : 0.93), (o_3, ∂ : 0.48), (o_4, ∂ : 0.86), (o_5, ∂ : 0.52), (o_6, ∂ : 0.78), (o_7, ∂ : 0.24), (o_8, ∂ : 0.70), (o_9, ∂ : 0.83), (o_{10}, ∂ : 0.90), (o_{11}, ∂ : 0.37), (o_{12}, ∂ : 0.05)가 구성되어 있다고 가정 할 때 FAS와 도메인 문서 객체와의 그룹화 관계는 다음과 같다.



(그림 5) FAS와 도메인 문서 객체와의 퍼지 관계

3.3.1 FAS : 0.6 이상을 만족하는 객체 그룹화

(그림 5)에서 $SRO(D_{omain}, G_j) \geq 0.6\text{-}cut_{(SRO)}$을 만족하는 문서 객체는 o_1, o_2, o_4, o_6, o_8, o_9, o_{10}이며 따라서 FAS : 0.6과 관련된 그룹화는 {o_1, o_2, o_6, o_9, o_{10}}이 되게 된다.

3.3.2 FAS : 0.7 이상을 만족하는 객체 그룹화

$SRO(D_{omain}, G_j) \geq 0.7\text{-}cut_{(SRO)}$을 만족하는 문서 객체는 o_1이며 이는 (그림 5)에서 FAS : 0.7과 퍼지 필터링을 수행하는 문서 객체는 o_1, o_3, o_5, o_7, o_{11}, o_{12}의 6개이지만 $SRO(D_{omain}, G_j) \geq 0.7\text{-}cut_{(SRO)}$을 만족하는 문서 객체는 o_1이다. 따라서 FAS : 0.6과 관련된 그룹화는 {o_1}이 되게 된다.

3.3.3 FAS : 0.8 이상을 만족하는 객체 그룹화

$SRO(D_{omain}, G_j) \geq 0.8\text{-}cut_{(SRO)}$을 만족하는 문서 객체는 o_2, o_4, o_9, o_{10}이며 따라서 FAS : 0.8과 관련된 그룹화는 {o_2, o_4, o_9, o_{10}}이 되게 된다.

4. 실험적 평가

이 절에서는 본 논문에서 제안된 퍼지 필터링 관련성 객체 그룹화의 성능을 분석하기 위해 실험적 평가를 수행한다. 실험 측정을 위해서 서버는 Windows-2000 서버 상에서 SQL 서버 7.0을 이용하였으며, 클라이언트는 windows2000의 Microsoft VisualC++6.0과 MFC를 이용하여 측정하였다. 측정데이터는 야후, 네이버, 라이코스 등의 검색엔진과 통계청 서버에서 "농림어업통계", "물가통계", "환경통계", "지역통계", "경기종합지수", "설비투자지수", "서비스업", "전자상거래", "정보산업" 분야와 관련된 텍스트, 이미지, 비디오프레임의 객체정보를 이용하였으며, 각각의 객체정보를 1600개의 문서정보로 분류하여 실험을 수행하였다. 실험을 위해 본 논문에서는 랜덤 키 방법(Random Key Method), OGM 방법, 제안된 방법으로 나누어 실행하며, 문서 구조화 방법의 OGM, BGM, RGM 기법에 있어서 비교적 나은 기법을 보이고 있는 OGM 기법과 실험 평가를 수행한다. 그리고 실험평가를 위해서 같은 클래스 정보를 구성하고 있는 객체 정보들은 같은 파일로 구성된다고 가정한다. 실험 평가를 위해 응용 도메인에서 임의의 객체들에 대한 검색시간으로

$$O_{searchtime} = O_T(1+1/N) + \frac{OT}{Doc} \times O_{Randomkey}$$

를 이용하며, 여기서 $O_{searchtime}$은 해당 문서 객체를 찾는데 걸리는 시간이다. 문서 객체를 검색하는데 이용되는 검색시간 연산자는 <표 1>과 같다.

〈표 1〉 검색시간 연산자

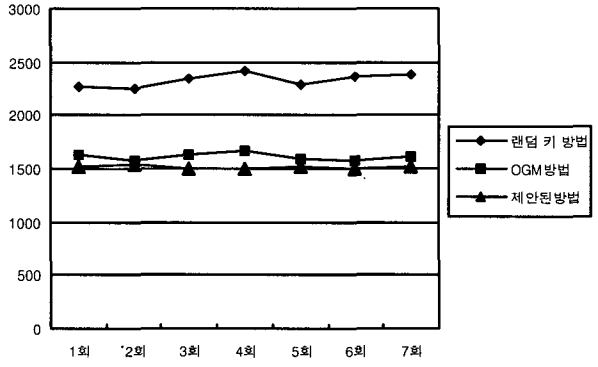
연 산 자	기 능 정 의
O_N	전체 객체정보들의 수
O_{time}	∂를 만족하는 문서 객체정보들을 찾는데 걸리는 시간
GO_{time}	그룹에서 FAS를 만족하는 문서 객체 정보를 찾는데 걸리는 시간
$DIST_{time}$	그룹에서 그룹, 디렉토리에서 디렉토리, 파일에서 그룹 및 디렉토리로 이동하는데 걸리는 시간

응용 도메인에서 ∂객체를 검색하는데 걸리는 시간은 $O_{time} = GO_{time} + DIST_{time} \times O_N$이며, FAS를 만족하는 문서 객체 정보를 찾는데 걸리는 시간은 $GO_{time} = x(O_i) + y(O_j)$이다.

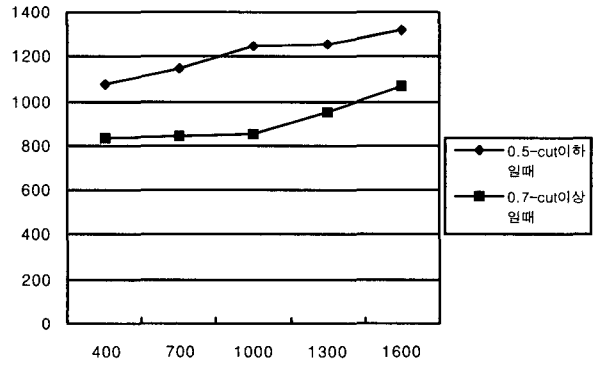
즉 $x(O_i)$는 $(D_{omain}, G_j) \geq \partial\text{-}cut_{(SRO)}$를 만족하는 객체 타입 정보를 찾는데 걸리는 총 평균시간을 말하며, $y(O_j)$는 객체정보 O_j가 FAS를 만족하는 객체 타입정보를 찾는데 걸리는 총 평균 시간을 말한다. 따라서 1600개의 객체정보를 문서별로 선정하여 7회를 반복 수행한 후 걸린 총 평균 검색시간은 (그림 6)과 같다.

(그림 6)에서 본 바와 같이 제안된 방법이 랜덤 키 방법과 OGM을 이용한 방법보다 해당 객체를 찾는데 걸리는 시간이 적게 걸리게 됨을 알 수 있다. 즉 랜덤 키 방법의 총 평균시간은 2329, OGM 방법의 총 평균 시간은 1614인데 반해 제안된 방법은 1514이다. 랜덤 키 방법, OGM 방법은 임의의 문서 객체 정보를 검색엔진 상에서 키워드를 이용하여 해당

객체를 탐색하는데 반해서 제안된 방법은 해당 객체를 FAS를 수행하여 그룹 안에서 해당 객체를 찾기 때문이다. 한편 OGM 방법은 관련된 문서들에 대해서 유사도를 수행한 관계로 랜덤 키 방법에 비해 검색 성능이 높게 나타남을 알 수 있다. 그리고 1,600개의 문서 객체정보들에 대해서 객체정보를 400, 700, 1,000, 1,300, 1,600개로 문서 정보들을 단계별로 분류한 후 ∂-cut이 0.5이하일 때와 0.7 이상일 때를 적용한 시간적 총합은 (그림 7)과 같다. (그림 7)에서 보는 것처럼 ∂-cut이 0.5-cut이하일 때는 문서 객체 정보들의 시간적 총합이 문서 객체 수의 증가에 따라 검색 시간이 계속 증가되고 있지만 제안된 방법은 ∂-cut이 0.8-cut 이상일 때 문서 객체 정보 수의 증가에 관계없이 검색시간이 줄어들었음을 알 수 있다.



(그림 6) 총 평균 검색 시간



(그림 7) α-cut이 0.5 이하일 때와 0.7 이상일 때의 검색 시간

따라서 제안된 방법은 관련된 문서 객체 그룹화에 있어서 검색성능이 우수함을 알 수 있으며, 멀티미디어 통계 문서정보에 대해서 문서 정보의 접근을 용이하게 지원해 줄 수 있는 장점을 가지게 된다.

5. 결론 및 향후 연구

본 논문에서는 웹 상에서 사용자가 요구하는 멀티미디어 통계 문서 정보를 객체 관련성에 따라 서비스하기 위해 퍼지 필터링 기반의 관련성을 이용하여 문서객체들을 그룹화하는 방법을 제안하였다. 일반적으로 응용 도메인에서 관련된 문서객체를 그룹화하기 위해서는 응용도메인의 문서 객체를 탐색해야 되며, 탐색 결과에 따라 그룹화 여부가 결정되게 된다. 본 논문에서는 그룹화를 위해 문서 객체를 프로파일 객체 타입으로 구성하였으며, 객체 프로파일에 따라 데이터 관련성을 결정하기 위해 데이터 관련성을 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$ 관계로 분류하였다. 또한 $RelCRO(D_{omain}, G_i)$와 $RelSRO(D_{omain}, G_i)$ 관계에서 관련된 객체를 필터링하기 위하여 $SRO(D_{omain}, G_i) \geq \partial\text{-}cut_{(SRO)}$를 제안하였다. 퍼지 필터링이 수행된 각 문서 객체에 대해서는 FAS를 수행하여 그룹화를 결정되게 되며, 본 논문에서는 FAS를 수행할 ∂-cut를 0.5-cut 이상을 만족하는 문서객체들에 대해서만 FAS를 수행하였다. 그 결과 1600개의 멀티미디어 문서 객체들에 대해서 실험을 수행한 결과 본 논문에서 제안된 방법의 성능이 우수함을 보였다. 향후 연구로는 본 논문에서 제안된 기법을 실제 멀티미디어 통계 문서 정보들을 미디어 타입별로 DB화하여 서비스할 수 있는 시스템 개발이 요구된다.

참 고 문 헌

[1] Andrea Rodriguez and Max J. Egenhofer, "Determing Semantic Similarity among Entity Classes from Different Ontologies," IEEE Transactions on Multimedia, Vol.15, No.2, pp.442-456, 2003.

[2] Chabane Djeraba, "Content-Based Multimedia Indexing and Retrieval," IEEE Multimedia, pp.18-22, 2002.

[3] Elina Megalou and Thanasis Hadzilacos, "Semantic Abstractions in the Multimedia Domain," IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.1, pp. 136-160, 2003.

[4] Elisa Bertino, Jianping Fan, Elena Ferrari, Mohand-Said Hacid, Ahmed K. Elmagarmid and Xingquan Zhu, "A Hierarchical Access Control Model for Video Database Systems," ACM Transactions on Information Systems, Vol.21, pp.155-191, 2003.

[5] Guojun Lu, "Techniques and Data Structure for Efficient Multimedia Retrieval Based on Similarity," IEEE Transactions on Multimedia, Vol.4, No.3, pp.372-384, 2002.

[6] Hsinchun chen, chris schuffels and Richard Orwig, "Internet Categorization and serch : self-organizing approach," Jounal of visual communication and image representation, Vol.7, No.1, pp.88-102, 1996.

[7] K. Bohms and T. C. Rakow, "Metadata for Multimedia Documents," SIGMOD Record, Vol.23, No.4, pp.21-26, 1994.

[8] Laszlo T. Koczy and T. Geodeon, "Information Retrieval by Fuzzy Reletions and Hierarchical Co-occurrence," Part I. TR97-01, Dept. of Info. Eng., School of Com. Sci. & Eng., UNSW, pp.1-18, 1997.

[9] Luis Gravano, Panagiotis Ipeirotis and Mehran Sahami, "QProber : A System for Automatic Classification of Hidden-Web Databases," ACM Transactions on Information Systems, Vol.21, pp.1-41, 2003.

[10] Pavel Calado, Edleno Moura and Ilmerio Silva, "Local versus Gloval link information in the Web," ACM Transactions on Information Systems, Vol.21, pp.42-63, 2003.

[11] Prasanna Ganesan, Hector Garcia-Molina and Jennifer Widom, "Exploiting Hierarchical Domain Structure to Compute Similarity," ACM Transactions on Information Systems, Vol.21, pp.64-93, 2003.

[12] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani and Liadan O'Callaghan, "Clustering Data Streams : Theory and Practice," IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.3, pp.515-525, 2003.

[13] 허문열, XLISP-STAT 객체 지향 통계언어, 자유아카데미, 1995.

이 종 득

e-mail : cdlee1008@iksan.ac.kr

1983년 전북대학교 전산통계학과(이학사)

1989년 전북대학교 대학원 전산통계학과 (이학석사)

1998년 전북대학교 대학원 전산통계학과 (이학박사)

1992년~2002년 서남대학교 컴퓨터정보통신학과 교수

2002년~현재 국립 익산대학 정보통신공학과 교수

관심분야 : 멀티미디어 통신, 임베디드 시스템, 무선인터넷, 무선 통신 등

김 대 경

e-mail : dkkim@chonbuk.ac.kr

1983년 전북대학교 전산통계학과(이학사)

1985년 동국대학교 통계학과(이학석사)

1995년 동국대학교 통계학과(이학박사)

1996년~1997년 국립 삼척대학교 교수

1997년~현재 전북대학교 통계정보학부 교수

관심분야 : 신뢰성, 표본이론, 멀티미디어통계