정보관리연구, vol.40, no.2 2009, pp.95-115

한글 저자명 군집화를 위한 계층적 기법 비교*

Exploration of Hierarchical Techniques for Clustering Korean Author Names

강 인 수*
In-Su Kang

차 례

초 록

저자식별은 학술문헌에 출현한 동명저자명들을 실세계의 서로 다른 사람들로 대응시키는 것이다. 이를 위해 임의의 동명저자명쌍의 유사도를 계산하고 이를 바탕으로 동명저자명 개체들을 군집화하는 단계를 거친다. 저자명의 군집화 기법으로 주로 계층적 군집법이 사용되었으나 다양한 계층적 군집법에 대한 비교 평가는 미흡했다. 이 연구는 다이스계수, 코사인유사도, 유클리디안 거리, 자카드계수, 피어슨 상관계수 등의 다양한 개체거리/유사도수식과 계층적 군집법들의 상관관계와 계층적 군집기법들의 한글 저자식별 성능에 대한 비교/분석을 다룬다.

키 워 드

저자식별, 계층적 군집법, 거리함수

* 본 논문은 2009학년도 경성대학교 학술연구비지원에 의한 것임.
** 경성대학교 컴퓨터정보학부 전임강사
(Full-time Lecturer, Kyungsung University, dbaisk@ks.ac.kr)
• 논문접수일자 : 2009년 5월 10일
• 게재확정일자 : 2009년 6월 23일

ABSTRACT

Author resolution is to disambiguate same-name author occurrences into real individuals. For this, pair-wise author similarities are computed for author name entities, and then clustering is performed. So far, many studies have employed hierarchical clustering techniques for author disambiguation. However, various hierarchical clustering methods have not been sufficiently investigated. This study covers an empirical evaluation and analysis of hierarchical clustering applied to Korean author resolution, using multiple distance functions such as Dice coefficient, Cosine similarity, Euclidean distance, Jaccard coefficient, Pearson correlation coefficient.

KEYWORDS

Author Disambiguation, Hierarchical Clustering, Distance Function

1. 서 론

학술문헌의 검색에서 저자명 검색은 동명이인의 존재로 인해 검색 정확률이 저하되는 어려움이 있다. 이 문제의 해법으로 제시된 저자명 중의성 해소(author name disambiguation), 혹은 저자식별(author resolution)은 같은 이름의 저자명 개체를 실세계의 서로 다른 사람들에 해당하는 식별자에 대응시키는 방법이다. 저자명 중의성 해소는 일반적으로 동명이인들 각각에 해당하는 동명저자명 개체들의 그룹을 만들기 위해 특정 거리/유사도 함수를 이용하여 모든 동명저자명 개체쌍의 거리/유사도를 계산한 다음 이를 바탕으로 군집화(clustering)를 적용한다. 군집화가 아닌 분류적 접근(Han et al. 2004)도 시도되었으나, 동명저자명 개체 집합에 대응하는 실세계 사람의 수가 몇 명인지 알려지지 않은 상황을 감안할 때 실세계 사람들에 대응하는 학습데이터의 존재를 가정해야 하는 분류기법(classification approach)의 적용은 적절치 못하다.

저자명 군집화를 위한 군집기법으로 주로 계층적 군집법(hierarchical clustering)이 시도되었다. 대표적 계층적 군집법으로 단일링크법(Single-linkage), 완전링크법(Complete-linkage), 대표링크법(Average-linkage), 워드법(Ward method)들이 있으며, 대표링크법은 네 가지 변이형이 존재한다. 이들 각 기법들은 군집 형성 과정이 달라 군집 결과에서도 차이를 보이는 것이 일반적이다. 예를 들어, 단일링크 군집법은 체인 모양으로 길게 연결된 군집들을 형성하고, 완전링크 군집법은 덩어리

모양의 응집된 군집들을 형성하는 특성이 있다(Sneath & Sokal 1973).

저자명 군집화의 기존 연구에서 일부 계층적 군집법에 대한 비교가 있었으나(Song et al. 2007; Tan et al. 2006), 그 범위가 제한적이었으며 저자명 군집화 문제에 대한 다양한 계층적 군집기법들에 대한 체계적 평가는 시도된 적이 없다. 자동 저자식별 시스템을 구축할 때 적절한 군집기법과 거리함수의 선택은 피할 수 없을 것이다. 또한 저자명 개체의 자질 표현을 결정하는 문제와 계층적 군집법을 사용할 경우 군집종료조건을 결정하는 문제도 해결되어야 할 것이다. 이 연구는 한글 저자식별 문제에서 다양한 계층적 군집법의 비교 평가를 통해 전술한 문제들을 해결하기 위한 실험적 자료를 제시하고자 한다.

일반적으로 군집화(clustering)는 먼저 군집 대상 개체 집합에 속한 각 개체를 자질벡터로 표현하고, 임의의 두 개체 사이에 계산되는 거리나 유사도를 바탕으로 군집기법을 적용하여 개체 군집들을 만드는 방식으로 동작한다. 이 연구에서는 군집 대상이 되는 저자명 개체의 자질벡터 표현을 위해 저자명이 출현한 논문의 제목, 공동저자명(들), 게재지, 출판년도 등의 기본 서지항목들을 사용하며, 평가의 현실성을 고려하여 저자명이 출현한 논문의 원문에서 추출되어야 하는 전자메일주소, 소속, 초록 등의 정보는 활용하지 않는다. 저자명 개체들에 대응하는 자질벡터들 사이의 거리/유사도 계산을 위해서는 다이스(Dice)계수, 코사인(Cosine) 유사도, 유클리디안(Euclidean)거리, 자카드(Jaccard)계수, 피어슨(Pearson) 상관계수 등의 다양한 거리 함수를 적용한다. 군집기법으로는 전술한 일곱 가지 계층적 군집법을 사용하며 거리함수와 군집기법의 결합 및 저자명 군집에 적합한 군집기법에 대한 비교 평가를 제시한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한다. 3장에서는 저자명 개체 쌍의 거리 함수들, 계층적 군집 기법들에 대해 기술한다. 4장과 5장에서는 저자식별을 위한 계층적 군집법의 실험 방법과 그 결과를 기술하고, 6장에서 결론을 맺는다.

2. 기존 연구

저자명 군집화는 전자도서관 검색서비스 향상을 위한 목적으로 본격적으로 연구되기 시작하였다(Alani et al. 2003; Han et al. 2003). 일반적인 저자명 군집화의 절차는 인명 변이형들[1]을 다루기 위한 인명매칭 단계를 거치며 이는 레코드 링키지 분야 연구(Bilenko et al. 2003; Elmagarmid et al. 2007)와 관련된다.

저자식별에 적용된 기존 군집법으로는 전통

1) "John F. Sowa", "J. F. Sowa", "J. Sowa", "Sowa, J.", "Sowa, J. F." 등과 같이 주로 영어 이름 표기에서 흔히 발견된다.

적인 계층적 군집법이 대부분이다. Song 등(2007)은 비슷한 수준의 좋은 클러스터링 성능을 보인 단일링크법, 완전링크법, 워드법 중 완전링크법을 선택했다고 논문에 적었다. Tan 등(2006)은 단일링크법, 완전링크법, 단순링크평균법(group average)의 저자식별 성능을 제시하면서 같은 군집으로 합병되어야 할 저자명 개체들 사이에 공유되는 자질들이 많지 않은 상황에서는 단일링크법이 우수함을 언급하였다. 강인수는 단일링크법을 한글 저자식별에 적용해 왔다(강인수 et al. 2008; 강인수 2008a; 강인수 2008b; Kang et al. 2009). Huang 등(2006)은 대용량 학술문헌에 대한 효율적 처리를 위해 개체들의 밀집도를 기반으로 군집을 형성하는 DBSCAN 클러스터링 기법을 저자식별에 적용하였다.

군집 알고리즘은 군집 대상 개체들의 개체거리 혹은 개체 유사도값을 기반으로 동작하므로 개체 거리함수에 영향을 받을 수 있다. Song 등(2007)은 유클리디언 거리함수를 사용하였고, Tan 등(2006)은 코사인유사도를 사용하였다. Huang 등(2006)은 자질유사도들을 벡터성분으로 갖는 자질벡터를 SVM의 입력으로 받아 계산된 분류 신뢰도값을 개체유사도로 사용하는 방식을 취했다. 여기서 자질유사도는 자질 타입에 따라 다른 유사도함수를 적용하여 얻어졌는데 전자메일과 URL은 편집거리를, 주소와 소속은 자카드 유사도를, 공동저자명 등의 이름에 대해서는 변이형 처리를 위해 Soft-TFIDF를 사용하였다. 강인수는 다이스유사도(강인수 2008b)와 이진거리함수(강인수 et al. 2008; 강인수 2008a)를 한글 저자식별에 적용하였다. 강인수는 또한 Huang 등(2006)의 연구에 영감을 얻어 분류적 관점의 교사학습법을 통한 개체유사도 계산을 위해 SVM을 포함한 다양한 기계학습기법을 적용한 저자식별 연구를 수행하였다(강인수 2008b).

정리하면 계층적 군집법의 경우 부분적으로 저자식별에 적용된 연구가 있으나 충분치 못하였고, 서로 다른 개체 거리함수들이 저자식별에 미치는 영향에 대해서는 그 기존 연구를 찾기 힘들다.

3. 계층적 군집법

계층적 군집법은 개체 집합을 입력으로 받아, 먼저 개체 집합에 속한 각 개체를 하나의 군집으로 만든 다음, 임의의 두 군집 사이에 계산되는 군집거리(군집유사도)를 바탕으로 거리가 가장 가까운(가장 유사한) 두 군집을 하나의 큰 군집으로 병합하는 절차를, 마지막 하나의 군집이 남거나 가장 가까운 두 군집 사이의 거리(유사도)가 특정한 군집병합 임계치 이상(이하)이 될 때까지 반복하는 군집법이다. 두 군집 사이에 정의되는 군집거리(군집유사도)는, 먼저(각 개체가 서로 다른 군집에 속하는) 임의의 두 개체쌍 사이의 개체거리(개체유사도)를 계산한 다음, 이 개체거리(들)를 입력

〈표 1〉 개체거리함수 ($\vec{u},\vec{v}$:개체의 벡터표현들, u_i : $\vec{u}$의 i번째 성분값, d: $\vec{u},\vec{v}$의 차원, $\delta(p)$:수식 p가 참이면 1 거짓이면 0, $n_{11}=\sum_{i=1}^{d}\delta(u_i>0 \text{and} v_i>0)$, $n_{01}=\sum_{i=1}^{d}\delta(u_i=0 \text{and} v_i>0)$, $n_{10}=\sum_{i=1}^{d}\delta(u_i>0 \text{and} v_i=0)$)

	개체거리(유사도)함수	개체거리수식
Binary distance	$B(\vec{u},\vec{v})=1-\delta(n_{11}>0)$	$B(\vec{u},\vec{v})$
Cosine similarity	$C(\vec{u},\vec{v})=\frac{\sum_{i=1}^{d}u_i v_i}{\sqrt[2]{\sum_{i=1}^{d}u_i^2}\sqrt[2]{\sum_{i=1}^{d}v_i^2}}$	$1-C(\vec{u},\vec{v})$
Dice coefficient	$D(\vec{u},\vec{v})=\frac{2n_{11}}{2n_{11}+n_{01}+n_{10}}$	$1-D(\vec{u},\vec{v})$
Euclidean distance	$E(\vec{u},\vec{v})=\sqrt[2]{\sum_{i=1}^{d}(u_i-v_i)^2}$	$\frac{E(\vec{u},\vec{v})}{d}$
Jaccard coefficient	$J(\vec{u},\vec{v})=\frac{n_{11}}{n_{11}+n_{01}+n_{10}}$	$1-J(\vec{u},\vec{v})$
Pearson correlation coefficient	$P(\vec{u},\vec{v})=\frac{\sum_{i=1}^{d}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt[2]{\sum_{i=1}^{d}(u_i-\bar{u})^2}\sqrt[2]{\sum_{i=1}^{d}(v_i-\bar{v})^2}}$	$\frac{1-P(\vec{u},\vec{v})}{2}$

으로 받는 군집거리함수에 의해 계산된다.

군집 대상 개체 사이의 개체거리(개체유사도) 계산을 위해 많은 개체거리 수식들이 제안되었다(Xu & Wunsch 2005). 대표적인 수식들을 정리하면 다음과 같다. 먼저 개체거리를 벡터공간 상에서 개체에 대응되는 두 벡터 사이의 거리로 보는 관점이 있으며, 유클리디언(Euclidean)거리, 시티블락(Cityblock)거리 등이 이에 속하며 모두 민코우스키(Minkowski)거리의 특수한 경우에 해당한다. 다음으로 벡터공간에서 두 벡터 사이의 각(angle)의 코사인값을 개체유사도로 보는 코사인 유사도(Cosine similarity)가 있다. 또한 두 개체의 벡터표현에서 대응하는 벡터성분값 쌍들의 상관관계(correlation)의 강도로 개체 유사도를 계산하는 관점이 있으며 피어슨(Pearson) 상관계수가 이에 속한다. 이진자질값을 갖는 개체표현에서 개체 유사도를 계산하는 수식에는 자카드계수(Jaccard coefficient)가 대표적이다. 이는 두 개체의 벡터표현에서 벡터성분값이 두 개체 모두 0인 벡터성분들은 제외하고 전체 벡터성분들 중 두 개체가 모두 0보다 큰 벡터성분값을 갖는 벡터성분들의 비로 계산된다. 자카드계수와 유사한 다이스계수(Dice coefficient)는 각 개체를 출현자질들의 집합 A, B로 표현했을 때

$2|A \cap B|/(|A|+|B|)$로 계산된다.

전술한 거리함수들 외에 이 연구에서는 이진거리(Binary distance)함수를 정의하여 사용한다. 이는 두 개체의 벡터표현에서 대응하는 벡터성분값이 두 개체 모두 0보다 큰 벡터성분이 하나 이상 발견되면 0의 거리를 그렇지 않은 경우 1의 거리를 할당하는 방식이다. 즉 두 개체를 표현하는 자질들 중 동시에 출현한 자질이 하나라도 있을 경우 0의 거리를 부여하는 방식이며 자질의 출현 여부가 개체의 식별에 결정적 영향을 미치는 경우에 유용할 것이다.

전술한 개체거리(개체유사도) 수식 중 이 연구에서는 이진거리, 코사인유사도, 다이스계수, 유클리디언거리, 자카드계수, 피어슨계수를 사용하며 이들은 〈표 1〉에 정리되어 있다.

계층적 군집법[2]에서 군집거리를 계산하는 방식은 단일링크(Single-linkage), 완전링크(Complete-linkage), 대표링크(Average-linkage)법이 있으며 이들은 두 군집의 모든 개체쌍 거리들의 최소값, 최대값, 대표값을 각각 군집 간 거리로 정의한다. 대표링크법은 대표값을 계산하는 방식의 차이에 따라 링크평균법, 군집중심법으로 세분된다. 링크평균법은 두 군집 사이의 모든 개체쌍 거리들의 평균을 계산하여 이를 군집거리로 사용하는 경우 단순링크평균법(UPGMA, group average)이 된다. 군집중심법은 각 군집에 속한 모든 개체 벡터들의 평균인 중심벡터(Centroid)를 계산하고 중심벡터들의 거리를 군집거리로 사용하는 경우 단순군집중심법(UPGMC)이 된다.

전술한 대표링크법들은 두 군집의 이전 군집병합 횟수나 군집크기(군집에 속한 개체의 개수) 등을 고려할 경우 다른 방식이 될 수 있다. 링크평균법은 원 개체거리에 대해 각 개체의 이전 군집병합 횟수에 지수적으로 반비례하는 가중치를 곱하여 얻어지는 가중개체거리들의 합을 군집거리로 사용하는 경우 가중링크평균법(WPGMA)이 된다. 그룹중심법은 한 군집의 중심벡터를 효율적으로 계산하기 위해 그 군집의 병합 전 이미 계산되어 있는 두 군집의 중심벡터들의 중심벡터(중앙)를 계산하는 경우 가중군집중심법(WPGMC, Median method)이 된다. 그러나 이 방법은 병합된 군집의 중심벡터가 병합 전 군집 크기가 큰 군집의 중심에서 작은 군집쪽으로 많이 이동하는 문제가 발생할 수 있다.

마지막 워드(Ward) 군집법(Ward 1963)은 병합된 이후 얻어진 새로운 군집의 분산과 병합되기 전 두 군집들의 분산의 합의 차(merging cost)를 군집간 거리로 사용한다.

4. 실험 방법

한글 저자명 저자식별의 성능 평가를 위해

2) 계층적 군집법에 대한 대부분의 내용들은 Sneath & Sokal(1973)에 기반한 것이다.

한국과학기술정보연구원에서 구축한 저자식별 평가셋[3](강인수 2008b)을 사용하였다. 실험에 사용된 평가셋은 1999년부터 2006년까지의 국내 정보기술 관련 주요 학술대회발표논문 7,677편에 출현한 2만614개의 저자명 개체들에 대해 실세계의 8,307명의 저자에 대응하는 저자식별자를 수작업으로 부여한 것이다. 2만614개의 저자명 개체(토큰) 중 5,164개의 서로 다른 저자명(타입)이 존재하며 5,164개의 동명저자그룹의 크기는 2부터 58까지 분포하고 있다.

3장에 기술한 개체 거리함수와 계층적 군집법 적용의 용이함을 위해 저자명 개체에 대한 자질 벡터표현을 만들 필요가 있다. 이를 위해 평가셋의 각 저자명 개체에 대해 그 저자명 개체가 출현한 논문의 공동저자명(들), 논문 제목, 게재지명[4], 게재연도에 출현하는 서로 다른 용어들을 서로 다른 벡터성분에 대응시키는 방식으로 자질 벡터표현을 만들었다. 가변개수의 다중 용어를 갖는 텍스트형 필드인 논문 제목의 경우는 형태소분석, 품사태깅을 거쳐 보통명사, 고유명사, 미등록어 명사로 태깅된 용어들을 추출하여 서로 다른 벡터성분에 대응시켰다.

저자명 개체의 벡터표현에서 각 벡터성분을 저자명 개체의 자질로 볼 수 있으며 벡터성분의 자질값은 이진가중치나 tfidf 가중치의 형태로 부여하여 저자식별에 미치는 영향을 각각 살펴보았다. 이진가중치 부여란 자질의 출현 시 1의 가중치가 그렇지 않은 경우 0의 가중치가 부여되는 방식이다. tfidf 가중치는 한 저자명 개체의 자질표현을 문헌(document)으로 고려하여, 특정 문헌에서 자질이 출현한 횟수(tf: term frequency)와 서로 다른 문헌들에서 자질이 출현한 횟수(df:document frequency)를 기반으로 다음 수식을 사용하여 계산된다(Manning et al. 2008).

$$w_{tfidf}(t) = (1+\log(tf(t)) \times \log(\frac{N}{df(t)})$$

〈식 1〉[5]

〈식 1〉에서 t는 자질, tf(t)는 자질 t의 tf값, df(t)는 자질 t의 df값이고, N은 저자명 개체의 전체 수이며 이 실험에서는 2만614이다.

저자식별 문제의 군집화 성능 평가를 위해 기존 연구들(Kang et al. 2009; Song et al. 2007)에서 사용된 pairwise-F1 지표를 사용하였다.

3) 이 평가셋은 한국과학기술정보연구원을 통해 배포됨(연락처: wksung@kisti.re.kr).

4) 한글 게재지명에 해당하는 영문약자를 용어로 사용하였다.

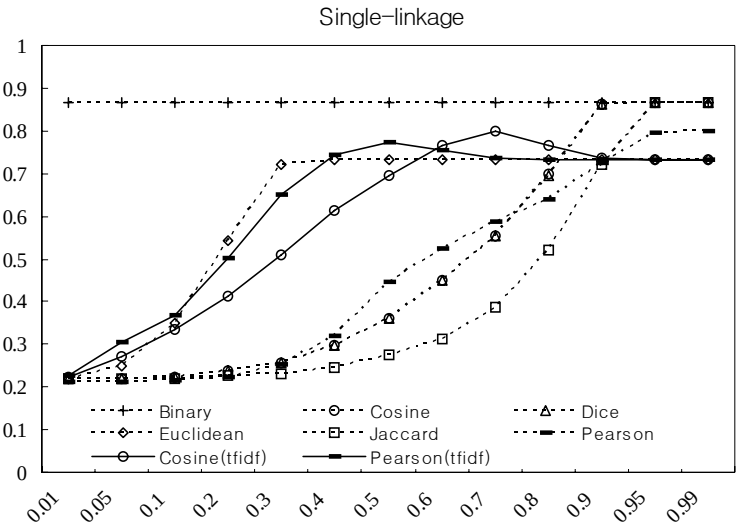
5) M개 논문의 저자명 개체의 전체 수가 N(N≥M)이라고 할 때, 각 저자명 개체의 자질표현에 대해 만들어지는 문헌의 전체 수는 N이다. 직관적으로는 저자식별과 관련하여 term frequency, document frequency보다 author feature frequency, author entity frequency라는 표현이 더 구체적일 수 있겠으나 tfidf 가중치를 부여하는 관점에서 보다 일반적인 정보검색 분야의 용어들을 사용한 것이다.

5. 실험 결과

군집기법과 거리함수의 결합 실험을 수행하기에 앞서 저자명 개체의 자질 추출을 위해 사용할 서지항목들을 결정하는 실험을 진행했다. 〈표 2〉는 그 결과를 보인 것으로 공동저자명(Coauthor: C), 논문 제목(Title: T), 게재지명(Publication: P), 게재연도(Year: Y)의 네 가지 서지항목들의 단일 및 다중 사용에 따른 저자식별 성능(F1)을 서로 다른 거리함수에 대해 제시하고 있다. 이 결과는 저자명 개체의 벡터표현에서 벡터성분값을 이진자질값으로 표현한 것이고 단일링크군집법을 적용한 것이다. 서지항목 중 공동저자명과 논문 제목의 이중 결합이 가장 좋은 성능을 보였으며 그 이상의 추가 서지항목의 사용은 성능을 저하시켰다. 유클리디언과 피어슨 거리함수를 제외한 나머지 거리함수들에서 대부분 동일한 성능을 보인 것은 최고 성능이 수렴된 데 기인한 것이며 수렴 이전의 군집병합 임계치의 변

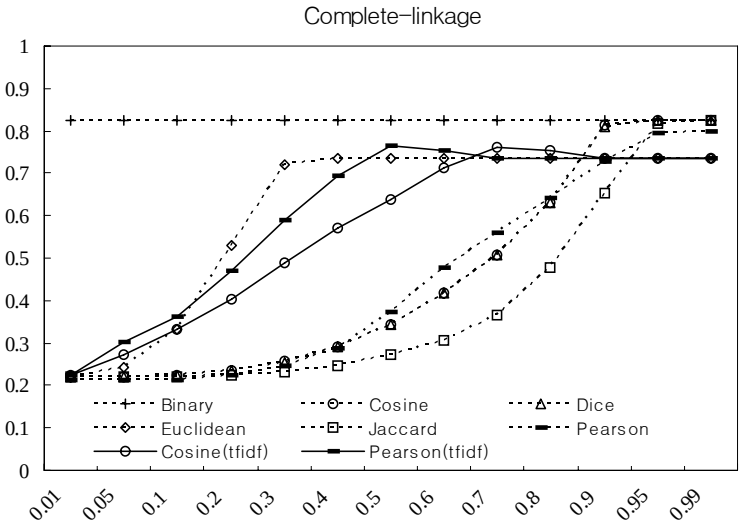
〈표 2〉 다중 자질과 저자식별 성능(이진자질값표현, Single-linkage, F1)

자질	Binary	Cosine	Dice	Euclidean	Jaccard	Pearson
C	0.8360	0.8360	0.8360	0.7302	0.8360	0.7873
C+T	0.8658	0.8658	0.8658	0.7335	0.8658	0.7984
C+T+P	0.8528	0.8528	0.8528	0.7335	0.8528	0.7912
C+T+P+Y	0.8148	0.8195	0.8148	0.7335	0.8208	0.7803

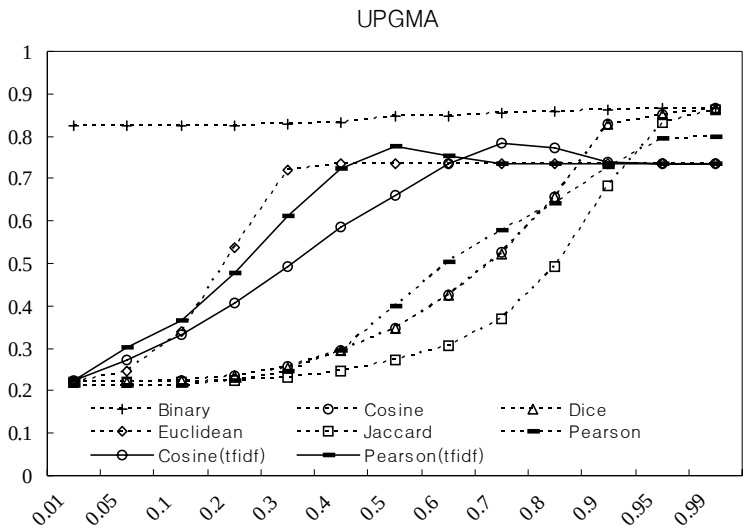


〈그림 1〉 서로 다른 거리함수를 사용한 Single-linkage 성능 (자질: C+T)

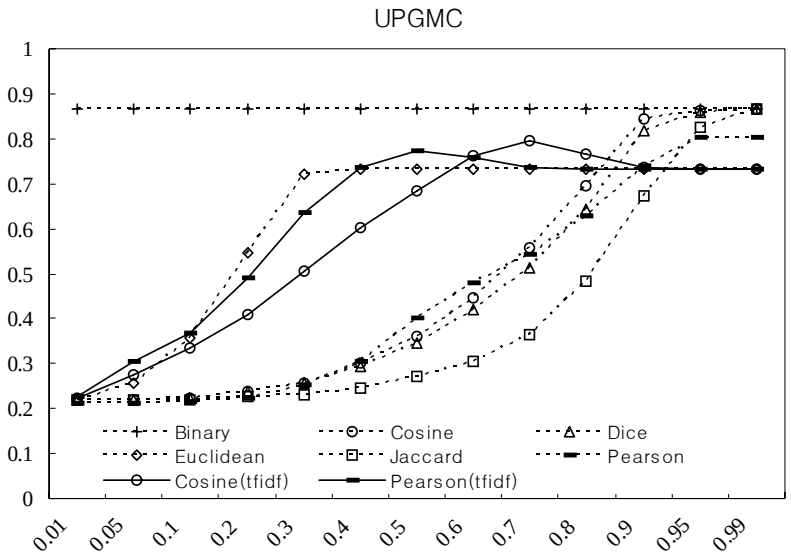
화에 따른 성능 추이는 상이하다(후술되는 단락 참조). 〈표 2〉를 통해 저자명 개체의 자질 추출을 위해 사용할 기본 서지항목으로 공동저자명과 논문제목 자질(C+T)을 결정하였고 이후 실험에서는 C+T, C+T+P, C+T+P+Y 각각을 자질로 사용한 경우의 저자식별 성능을 제시할 것이다.



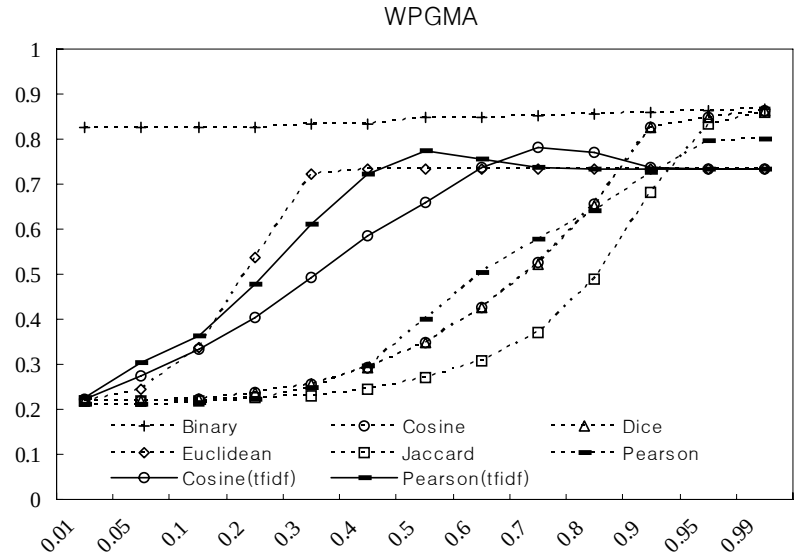
〈그림 2〉 서로 다른 거리함수를 사용한 Complete-linkage 성능 (자질: C+T)



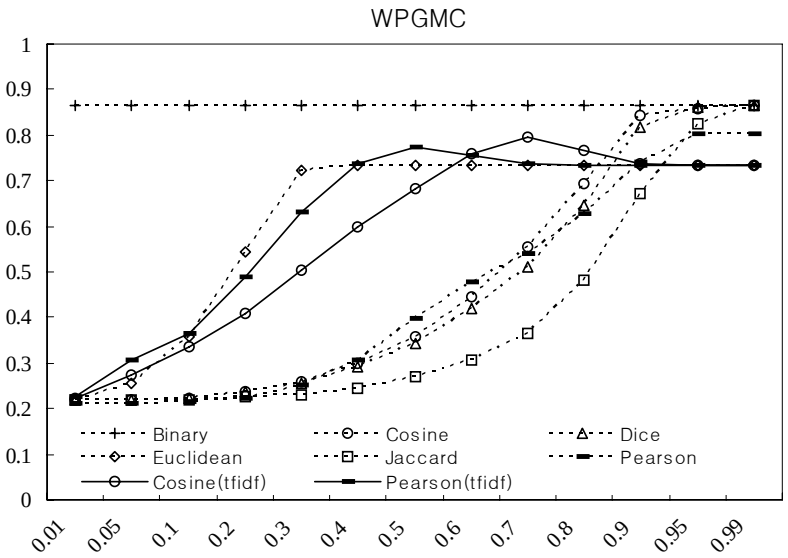
〈그림 3〉 서로 다른 거리함수를 사용한 UPGMA 성능 (자질: C+T)



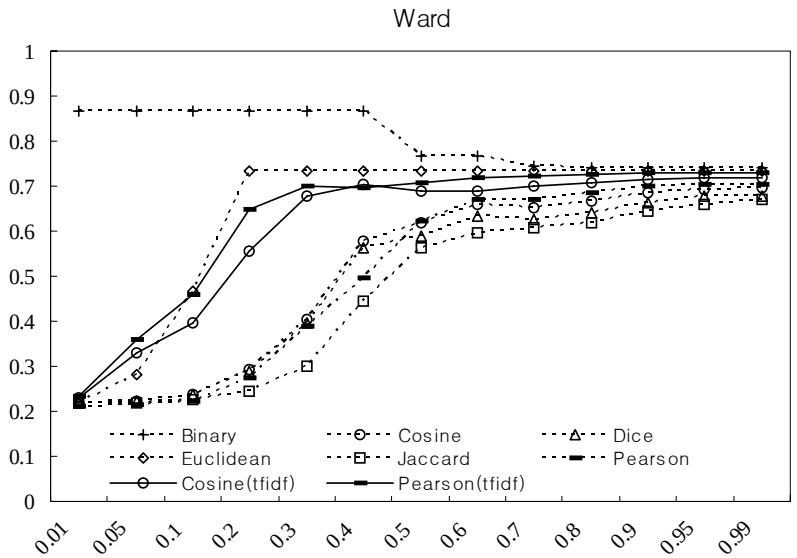
〈그림 4〉 서로 다른 거리함수를 사용한 UPGMC 성능 (자질: C+T)



〈그림 5〉 서로 다른 거리함수를 사용한 WPGMA 성능 (자질: C+T)



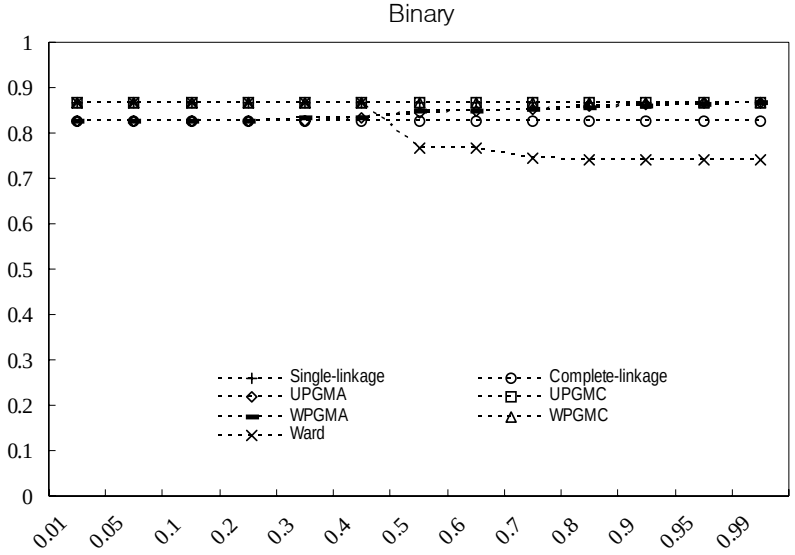
〈그림 6〉 서로 다른 거리함수를 사용한 WPGMC 성능 (자질: C+T)



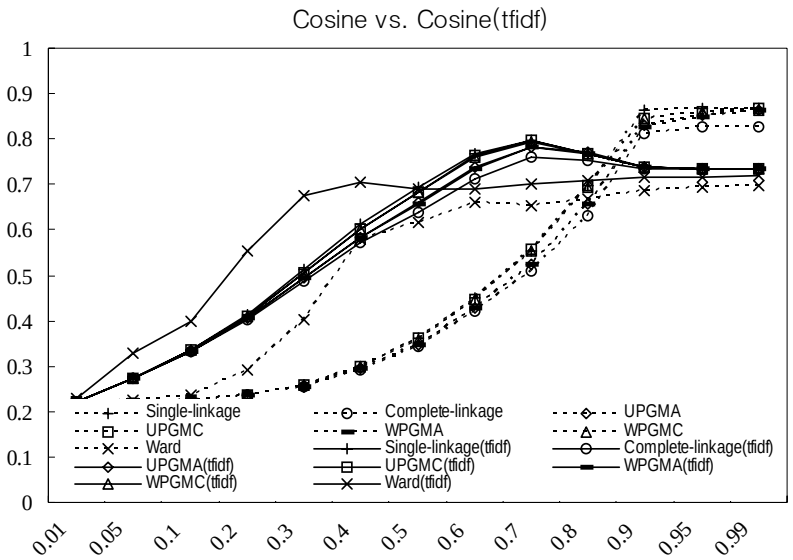
〈그림 7〉 서로 다른 거리함수를 사용한 Ward 성능 (자질: C+T)

〈그림 1〉~〈그림 7〉은 공동저자명과 논문 제목 자질(C+T)을 사용하여 각 계층적 군집법에 대해 서로 다른 거리함수를 사용한 경우의 저자식별 성능(F1)을 군집병합 임계치의 변화(0.01, 0.05, 1, 2, …, 0.9, 0.95, 0.99)에 따라 보인 것이다. 모든 군집기법들이 동일 군집기법 내에서 서로 다른 거리함수를 사용한 경우 군집병합 임계치의 변화에 따라 저자식별 성능에서 유의미한 차이를 보였다. 이는 저자식별 문제에서 거리함수 선택의 중요성을 보이는 실험 결과이다. 가장 좋은 성능을 보인 이진 거리함수는 군집병합 임계치의 변화에 거의 무관한 특성을 보였는데 이는 거리값을 0과 1 중 하나의 값으로만 계산하는 이진 거리함수의 특성에 기인한 결과이다.

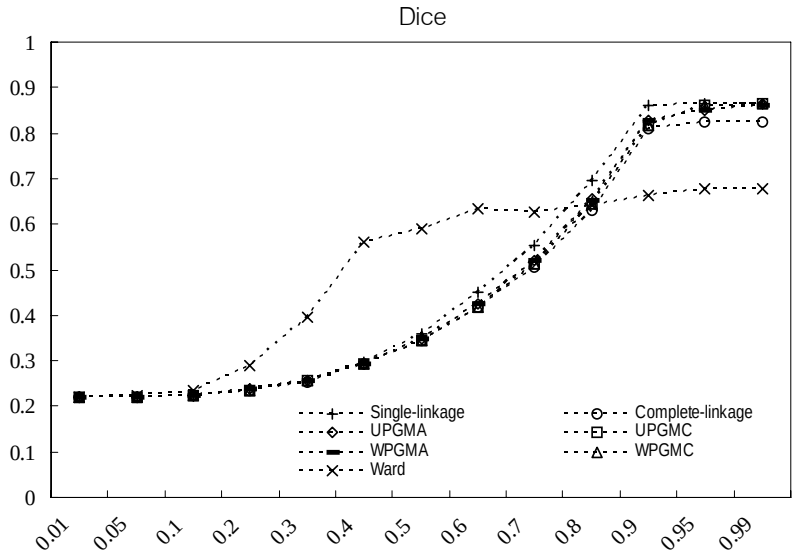
또한 〈그림 1〉~〈그림 7〉에서는 자질값 표현에서 출현 여부만을 고려하는 이진표현(점선)과 자질값(예: 공동저자명이나 논문 제목에 출현한 용어)의 TF와 IDF를 동시에 고려하는 tfidf 표현(실선)을 코사인, 피어슨 거리함수에 각각 적용한 저자식별 성능을 비교하고 있다. 논문의 기본 서지항목들을 하나의 문서로 고려하는 현재의 실험집합의 경우 거의 모든 용어의 TF가 1에 해당하므로 이 실험의 tfidf 표현은 IDF 표현이라 해도 크게 틀림이 없다. 모든 군집기법에서 코사인과 피어슨 거리함수에서 tfidf 표현이 이진표현에 비해 군집병합 임계치의 변화에 강인한 결과를 보였으나 최고 성능에서는 이진표현이 우수하였다. 이에 대한 분석은 후행 단락들에서 다루어진다.



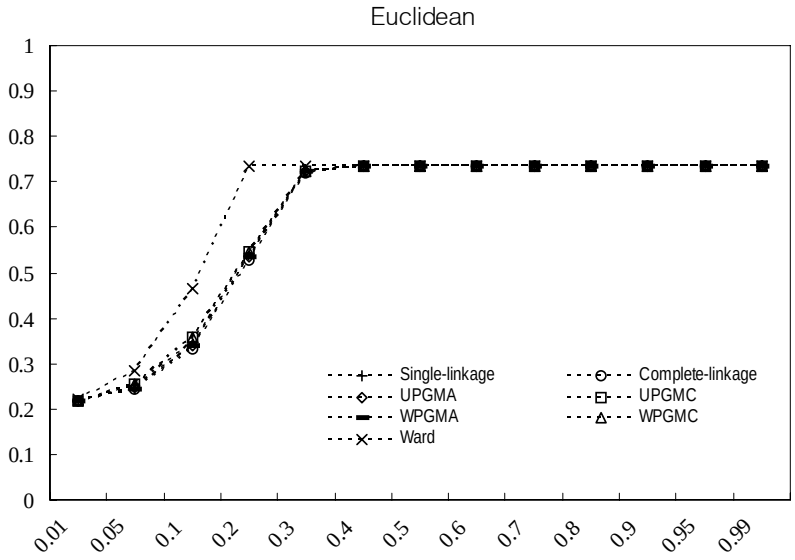
〈그림 8〉 Binary 거리함수에 대한 계층적 군집법 성능 (자질: C+T)



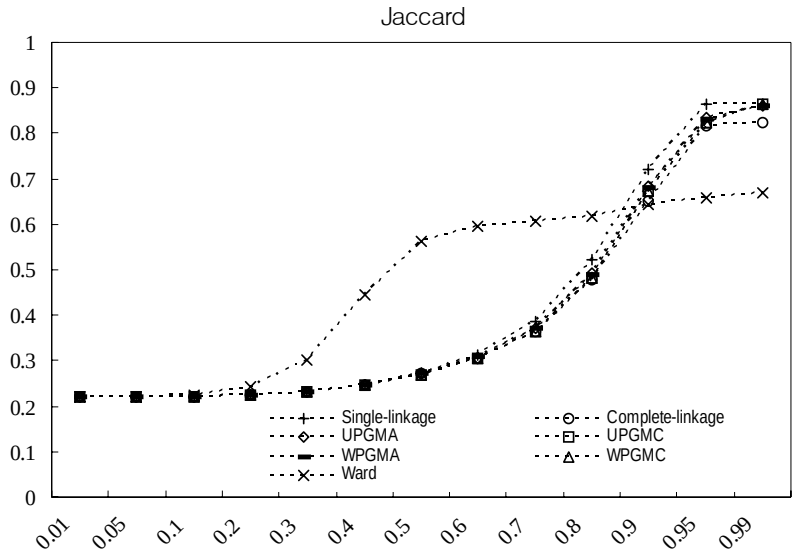
〈그림 9〉 Cosine 거리함수에 대한 계층적 군집법 성능 (자질: C+T)



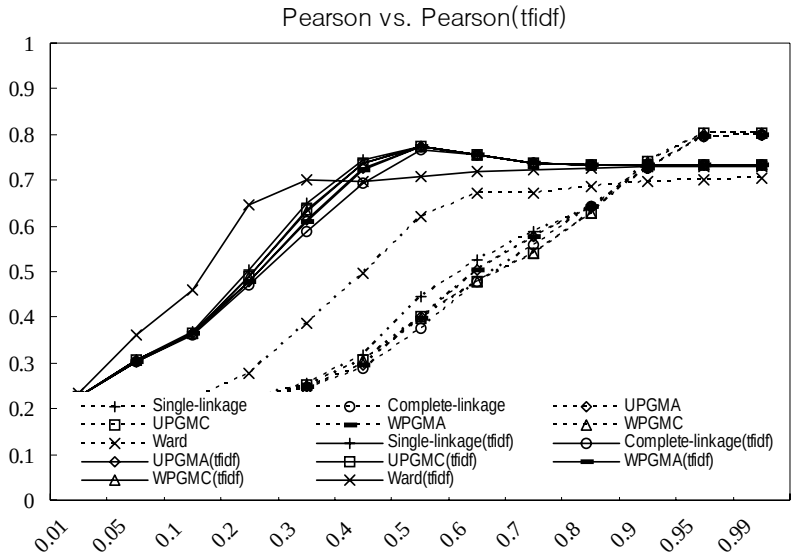
〈그림 10〉 Dice 거리함수에 대한 계층적 군집법 성능 (자질: C+T)



〈그림 11〉 Euclidean 거리함수에 대한 계층적 군집법 성능 (자질: C+T)



〈그림 12〉 Jaccard 거리함수에 대한 계층적 군집법 성능 (자질: C+T)



〈그림 13〉 Pearson 거리함수에 대한 계층적 군집법 성능 (자질: C+T)

〈그림 8〉~〈그림 13〉의 실험 결과는 공동저자명과 논문제목 자질(C+T)을 사용하여 같은 거리함수를 서로 다른 군집기법에 적용했을 경우의 저자식별 성능을 군집병합 임계치의 변화와 함께 보인 것이며, 코사인, 피어슨 거리함수의 경우는 각 군집기법에 대해 이진자질값 표현(점선)과 tfidf 자질값 표현(실선)을 동시에 나타내었다. 〈그림 8〉~〈그림 13〉에서 워드법을 제외한 모든 계층형 군집기법들은 비슷한 성능 추이를 보임으로써 저자식별 문제에서 계층형 군집기법들은 고유한 특성을 드러내지 못했다. 이는 실험에 사용된 평가셋의 경우 군집병합의 근거가 되는 군집 간 유사도 계산에서 서로 다른 두 군집에 속한 개체들의 집단적 특성의 사용이 저자식별 절차에 큰 영향을 미치지 않음을 의미한다. 집단적 특성의 사용이란 대표링크법에서 같은 군집 내에 속한 개체들의 대표적 개체 표현인 중심벡터(centroid)를 계산한다든가 완전링크법에서 서로 다른 군집에 속한 개체 집합 간에 정의되는 모든 개체쌍들이 최소 유사도 수준을 만족하는지를 검사하는 것 등을 뜻한다.

성능의 차이가 크지는 않으나 군집법 중 단일링크법이 가장 좋은 성능을 보였고 완전링크법이 워드기법을 제외하고 거의 모든 거리함수에서 가장 낮은 성능을 보였으며 대표링크법들은 단일링크법과 완전링크법의 중간 성능을 나타냈다. 이는 현재 실험집합에서의 저자명 자질표현이 대표링크법이나 완전링크법의 집단적 특성을 발현하기에 충분하지 못한

이유에 기인한 결과이다. 예를 들어 완전링크법은 개체 표현 공간 내에서 덩어리 모양으로 응집된 형태의 군집 특성을 보이는 개체들의 군집화 문제에 적합한 것으로 알려져 있다(Sneath & Sokal 1973). 저자식별의 대상이 되는 저자명들의 군집 특성이 덩어리 모양을 형성하지 않는다고 말하는 것은 부적절해 보인다. 오히려 저자명 표현을 위해 선별된 자질들이 한 저자의 신원을 식별하기에 충분하고 각 자질값들을 누락 없이 획득하는 것이 보장된다면, 동일 저자에 해당하는 저자명 표현들은 개체 표현 공간 내에 응집되어 있을 확률이 클 것이므로 저자식별의 경우에도 완전링크법은 적절한 군집법으로 기능할 수 있을 것이다. 이와 관련하여 분류의 관점에서 커널 함수를 통해 개체의 연산이 발생하는 자질 공간을 변경하는 SVM기법을 완전링크법과 연계하는 것은 향후 흥미로운 연구 주제가 될 것이다.

〈그림 8〉~〈그림 13〉에서 tfidf 자질값 표현을 사용한 경우를 제외하고 모든 거리함수에 대해 군집병합 임계치가 1에 가까울수록 저자식별 성능이 계속 상승하였는데 이는 일반적인 직관과 상치되는 부분이다. 즉 군집 간 거리가 크다는 것은 군집 간 유사도가 작다는 것이므로 결국 이는 유사도가 거의 0에 가까운 군집들을 병합할수록 저자식별에서 보다 유리한 결과를 만든다는 논리가 되기 때문이다. 이진 거리함수가 보인 결과도 이들과 다르다 할 수 없다. 결국 임의의 두 저자명 표현이 이진 자질값(binary representation)들의 벡터로 주어졌을 때 동시에 출현한 자질이 하나라도 있기만 하면 두 저자명은 같은 저자로 군집화하는 것이 현재의 실험 집합에서는 더 좋은 저자식별 성능을 보인다고 할 수 있다.

전술한 비직관적 결과는 다음의 두 가지 가정에 근거하고 있다고 생각된다. 첫째는 두 저자명의 이진 자질값 표현들에서 공유되는 자질(예: 공동저자명이나 논문 제목의 용어)의 출현(자질값 1에 해당)은 굉장히 드문 사건이어서 이러한 사건의 발생은 두 저자명 표현들이 실세계의 동명이인일 것이라는 가설을 기각하고도 남을 만큼 충분히 놀랍다는 것이다. 둘째는 현재 실험집합에서 저자식별 중의성이 그리 크지 않아(즉 대부분의 동명저자명들이 실세계의 한 사람에 대응함으로 인해), 실제로 무의미한 자질(예를 들어 논문 제목에 출현하는 '연구', '향상', '분석' 등의 용어[6])이라 하더라도 두 저자명 표현의 유사도를 0보다 크게 만들 수만 있으면 그러한 자질의 공유가 저자식별에 도움이 된다는 것이다.

두 가정 모두 자질로 표현되는 용어의 구체성(specificity) 정도와 관련이 있다. 첫째 가정에서는 저자명 표현에 사용된 용어들이 상당한 구체성을 갖고 있어야 할 것이다. 왜냐하면 보편적이거나 불용어 성격의 용어들은 두 저자명 표현에서 공유된다 하더라도 이를 발

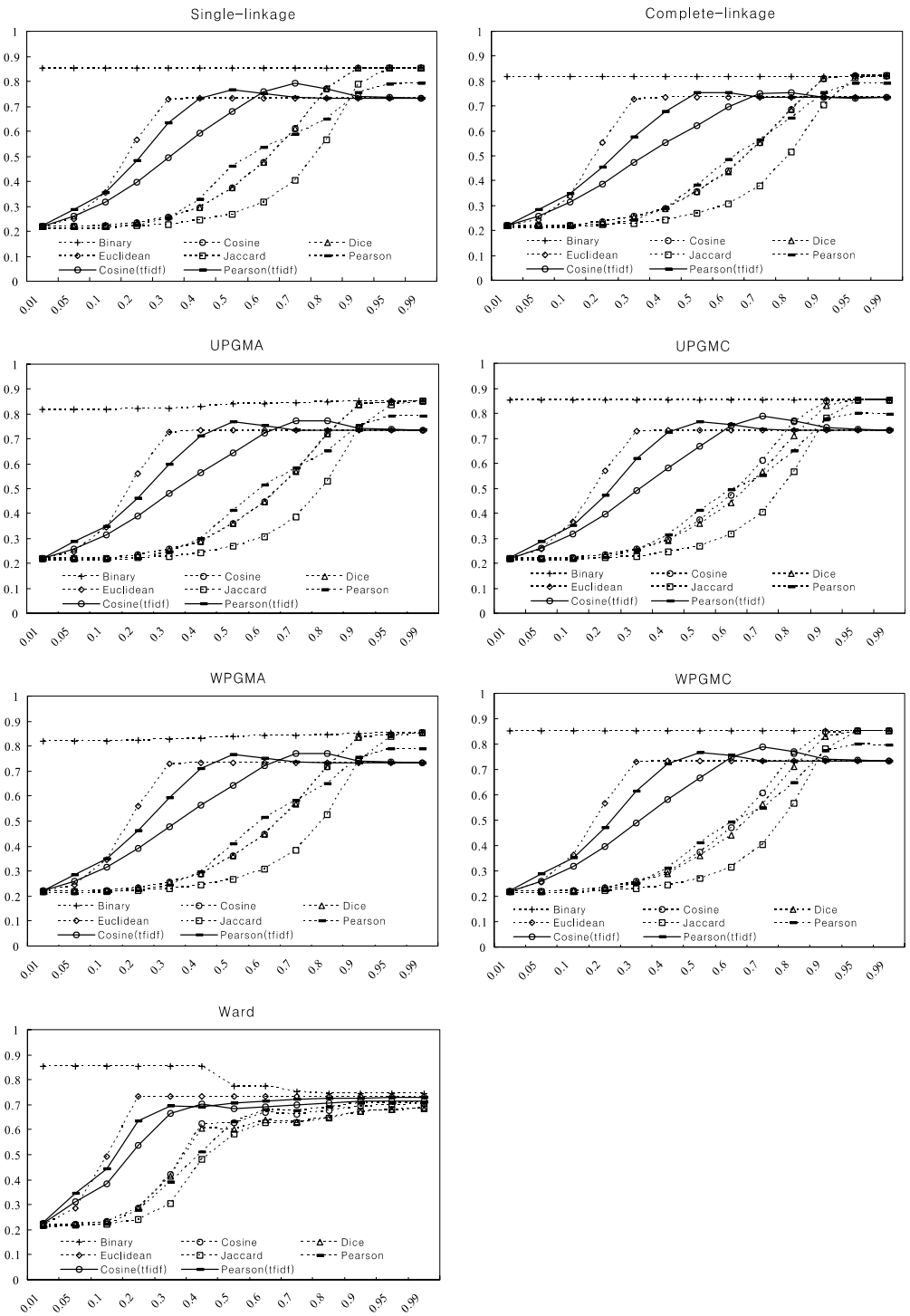
6) 컴퓨터 분야의 동명저자명 표현이 가질 수 있는 '컴퓨터', '정보' 등의 용어도 무의미한 자질에 해당될 수 있다.

생하기 힘든 사건으로 고려하기 어렵기 때문이다. 둘째 가정은 현재의 실험 결과를 설명하는 한 수단이 될 수 있을지 모르나 일반적인 저자식별 문제에 적용하기에 적절치 않다. 현재 실험집합에서도 〈표 2〉에 보인 것처럼 공동저자명이나 논문 제목 용어에 비해 구체성이 낮다고 생각되는 게재지명과 게재연도 자질을 추가적으로 사용한 경우 성능이 저하되는 결과를 보였기 때문이다. 즉 두 저자명 표현에서 게재지명과 게재연도 자질의 공유로 인한 두 저자명의 병합은 저자식별에 부정적 영향을 미친 것으로 해석될 수 있다.

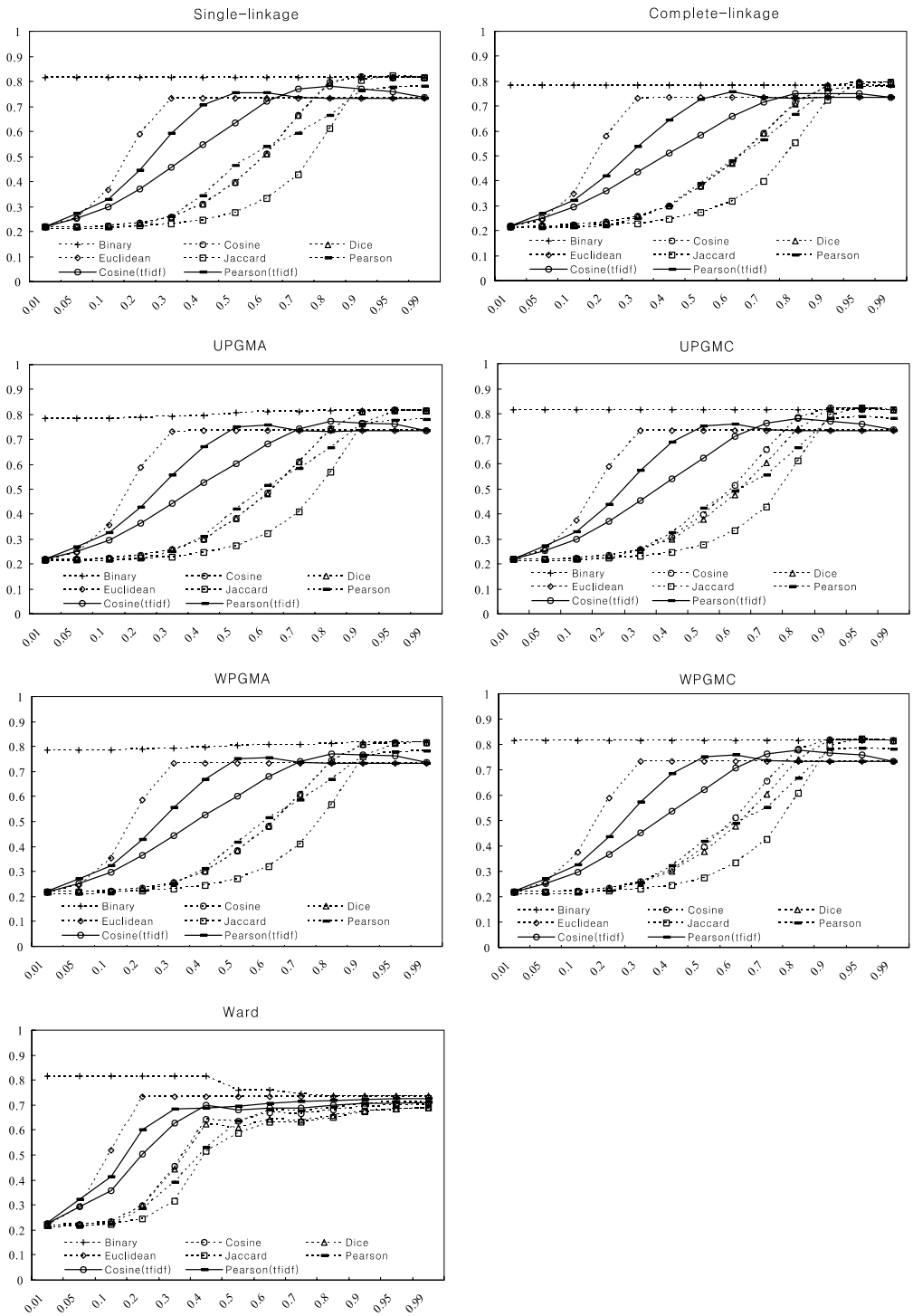
이제 둘째 가정을 배제하고 전술한 첫째 가정으로 돌아가면, 첫째 가정의 경우 공유되는 자질의 구체성 혹은 비보편성(rareness) 정도에 따라 공유 자질을 갖는 두 저자명의 병합 판단이 달라지는 복잡성을 갖는다. 〈그림 9〉, 〈그림 13〉의 코사인, 피어슨 거리함수에서 tfidf 자질값 표현(실선)은 자질의 비보편성을 IDF로 모델링한 경우로 볼 수 있다. 〈그림 9〉, 〈그림 13〉에서 tfidf 기반의 코사인과 피어슨 거리함수는 군집병합 임계치 0.5~0.7까지 성능 상승을 보인 이후 완만한 성능 감소를 보였다. 즉 tfidf 자질값 표현에서는, 이진 자질값 표현과 달리, 보편성이 큰 자질(들)의 공유로 인해 그 자질(들)을 공유하는 두 저자명을 실세계 동일인으로 병합할 위험을 피할 장치가 어느 정도 제공된다고 볼 수 있다. 그러나 최고 성능의 경우는 tfidf 표현보다 이진 자질값 표현이 우수하였는데 이는 현재 실험집합 내에서 얻어진 tfidf 값의 부정확함에 기인하는 것으로 추측된다. 향후 웹이나 대용량 문헌데이터베이스로부터 자질의 비보편성을 모델링하는 추가 연구가 필요할 것이다.

〈그림 1〉~〈그림 13〉은 저자식별 자질 중 공동저자명과 논문제목 자질(C+T)을 사용한 저자식별의 성능을 보인 것이다. 그러나 공동저자명, 논문제목 자질에 게재지명, 게재연도 자질을 추가하는 경우 저자명 개체의 자질 표현이 보다 풍부해질 수 있어 저자식별 성능의 변화를 가져올 수도 있다. 이의 확인을 위해 C+T+P, C+T+P+Y를 저자식별 자질로 사용한 경우 각각에 대해 저자식별 실험을 진행하였고 〈그림 14〉, 〈그림 15〉는 그 결과이다.

〈그림 14〉, 〈그림 15〉에서 알 수 있듯이 게재지명, 게재연도 자질의 추가 사용은 C+T 자질만 사용한 경우의 저자식별 성능과 큰 차이를 보이지 않았으나 전체적으로 성능을 저하시켰다. 이 결과는 저자식별 자질의 추가가 저자식별에 부정적 영향을 미친다는 것을 의미하는 것은 아니며, 현재 실험집합에서 C+T 자질에 대해 제시한 연구 결과들이 추가 자질들을 사용한 경우에도 유의미함을 보인 것이다. 저자식별 문제에 있어 자질의 종류와 수의 증가가 군집 성능에 미치는 영향에 대해서는 네 개 자질이 사용된 현재의 실험집합으로 평가하기에는 어려움이 있으며 향후 논문 원문이나 저자의 홈페이지로부터 얻어지는 자질들을 포함하는 보다 큰 자질 집합에서 이에 대한 평가가 진행되어야 한다.



〈그림 14〉 서로 다른 거리함수에 대한 계층적 군집법 성능 (자질: C+T+P)



〈그림 15〉 서로 다른 거리함수에 대한 계층적 군집법 성능 (자질: C+T+P+Y)

6. 결 론

이 연구는 학술문헌에 출현한 동명저자명을 실세계의 같은 사람에 해당하는 그룹으로 군집화하는 저자식별 문제에서 계층적 군집화 기법과 개체 거리수식의 상관관계에 대한 실험적 결과를 제시하였다.

같은 개체 거리함수를 사용할 경우 서로 다른 계층적 군집법들은 저자식별 성능에서 뚜렷한 차이를 보이지 못하면서 단일링크법, 대표링크법, 완전링크법, 워드법의 성능 순서를 보였으며 이는 저자식별 자질의 불충분성을 시사하는 실험결과로 해석된다. 같은 계층적 군집법을 사용할 경우 서로 다른 개체 거리함수들은 군집병합 임계치의 변화에 따라 저자식별 성능의 차이가 적지 않음을 보였으며 이를 통해 저자식별에서 적절한 개체 거리함수 선택이 중요함을 알 수 있었다.

또한 이 연구를 통해 저자명 개체에 대해 이진자질값 벡터표현을 사용한 경우 군집 알고리즘의 종료 조건에 해당하는 군집병합임계치 범위 결정에서 어려움이 발생함을 알 수 있었으며, 자질의 희소성(rareness)을 고려하는 IDF 자질값 표현이 전술한 문제에 대한 해법이 될 수 있는 가능성을 보였다. 향후 저자명 개체의 자질값 표현과 관련하여 자질의 비보편성을 모델링하는 기법들에 대한 추가 연구가 요구된다.

참고문헌

강인수, 이승우, 정한민, 김평, 구희관, 이미경, 성원경, 박동인. 2008. 저자식별을 위한 자질 비교. 『한국콘텐츠학회논문지』, 8(2): 41-47.

강인수. 2008a. 저자식별을 위한 전자메일의 추출 및 활용. 『한국콘텐츠학회논문지』, 8(6): 261-268.

강인수. 2008b. 한글 저자명 중의성 해소를 위한 기계학습기법의 적용. 『한국정보관리학회지』, 25(3): 27-39.

Alani, H., Dasmahapatra, S., O'Hara, K., & Shadbolt, N. 2003. "Identifying communities of practice through ontology network analysis." *IEEE Intelligent Systems*, 18(2): 18-25.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. and Fienberg, S. 2003. "Adaptive name matching in information integration." *IEEE Intelligent Systems*, 18(5): 16-23.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. 2007. "Duplicate record detection: A survey." *IEEE Transactions on Knowledge and Data Engineering*, 19(1): 1-16.

Han, H., Giles, C. L., and Zha, H. 2003. "A model-based k-means algorithm for name disambiguation." *Proceedings*

of semantic web technologies for searching and retrieving scientific data. October 20, Florida, USA.

Han, H., Giles, C. L., Zha, H., Li, C., and Tsioutsiouliklis, K. 2004. "Two supervised learning approaches for name disambiguation in author citations." *Proceedings of the ACM/IEEE joint conference on digital libraries(JCDL)*, 2004: 296-305.

Huang, J., Ertekin, S., and Giles, C.L. 2006. "Efficient name disambiguation for large scale databases." *Proceedings of PKDD-2006*, 2006: 536-544.

Kang, I.S., Na, S.H., Lee, S.W., Jung, H.M., Kim, P., Sung, W.K., and Lee, J.H. 2009. "On co-authorship for author disambiguation." *Information Processing and Management*, 45(1): 84-97.

Manning, C. D., Raghavan, P. and Schutze, H. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Sneath P. A. and Sokal R. R. 1973. *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman and Company.

Song, Y., Huang, J., Councill, I., Li, J., and Giles, C. L. 2007. "Efficient topic-based unsupervised name disambiguation." *Proceedings of the ACM/IEEE joint conference on digital libraries(JCDL)*, 2007: 342-351.

Tan, Y. F., Kan, M. Y., and Lee, D. W. 2006. "Search engine driven author disambiguation." *Proceedings of the ACM/IEEE joint conference on digital libraries (JCDL)*, 2006: 314-315.

Ward, J. H. 1963. "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association*, 58(301): 236-244.

Xu, R., and Wunsch, D. 2005. "Survey of clustering algorithms." *IEEE Transactions on Neural Network*, 16(3): 645-678.