www.calsec.or.kr
http://dx.doi.org/10.7838/jsebs.2012.17.4.209

대수적 특성을 고려한 벡터 유사도 측정 함수의 고찰

Survey on Vector Similarity Measures : Focusing on Algebraic Characteristics

이동주(Dongjoo Lee)*, 심준호(Junho Shim)**

초 록

전자 상거래 시스템 환경에서 상품, 상품평, 사용자 특성 등은 주요한 정보 객체이다. 벡터는 객체의 표현기법으로 널리 사용되고 있다. 전자 상거래 데이터 객체들은 벡터로서 모델되어 각 특질에 해당하는 차원의 숫자 값으로 표현될 수 있다. 전자 상거래의 특성상 이러한 객체들은 방대한 분량이 되고 있고, 이중 여러 객체들은 실제로 같거나 유사한 객체일 수 있다. 따라서 객체간 유사도 측정은 전자상거래 시스템에서 중요한 역할을 한다. 본 논문에서는 벡터 객체에서 사용되는 대표적인 유사도 측정 함수들을 고찰한다. 유사 함수들은 각각의 대수적 특성을 가지고 있고 서로 연결된 특성을 보인다. 이러한 특성을 분석하고 또한 유사 함수들을 분류해 본다. 이러한 과정은 표준 벡터 유사도 함수가 가져야 할 대수적 특성을 제시해준다.

ABSTRACT

Objects such as products, product reviews, and user profiles are important in e-commerce domain. Vector is one of the most widely used object representation scheme. Information of e-commerce objects may be modeled by vectors in which the featured values are assigned to various dimensions. E-commerce objects are in general quantitatively large while some are similar or even same in reality. It plays, therefore, an important role to measure the similarity between objects. In this paper, we survey the state-of-the -art vector similarity measures. Similarity measures are analyzed to feature the algebraic characteristics and relationship of those, and upon which we classify the related measures accordingly. We then present such features that standard vector similarity measures should convey.

키워드 : 유사도, 벡터, 근접성, 거리
Similarity, Vector, Proximity, Distance

본 연구는 숙명여자대학교 2011년 교비연구로 수행되었음.
* DMC R&D Center, Samsung Electronics Co
** Corresponding Author, Division of Computer Science, Sookmyung Women's University (E-mail : jshim@sookmyung.ac.kr)
2012년 10월 31일 접수, 2012년 11월 19일 심사완료 후 2012년 11월 20일 게재확정.

1. 서 론

웹과 인터넷의 발달로 인해 전자 상거래 시스템 환경에서 다루는 정보 객체의 양은 증가되고 있다. 상품에 대한 정보, 상품에 대한 사용자의 의견, 사용자에 대한 정보 및 특성 등은 전자 상거래에서 주요한 정보 객체라 할 수 있다. 정보 객체를 기계적으로 다루기 위해서는 객체를 표현하는 기법이 필요하며, 이 중 벡터는 객체의 표현기법으로 널리 사용되고 있다[3, 5, 7, 10].

예를 들어 웹상의 전자 상거래 상품에 대한 사용자의 리뷰 문서에서 필요한 의견을 자동으로 추출하고 분석하는 오피니언 마이닝(opinion mining) 분야에서, 사용자의 의견 정보는 대개 벡터로 모델된다[12].

예를 들어 사용자의 상품에 대한 의견 정보는 의견 표현의 대상이 되는 상품(o_i), 상품의 세부 특징(f_j), 의견 표현 표준 어휘(e_k), 사용자(u_m), 의견이 작성된 시간(t_n)과 위치(p_o) 등을 포함하는 벡터 $<o_i, f_j, e_k, u_m, t_n, p_o>$로서 표현될 수 있다.

웹상의 다양한 정보 객체들 가운데는 원래 비슷한 객체이었거나 같은 객체이었던 객체들이 중복되게 나타날 수 있다. 정보 객체가 벡터로서 표현된 경우, 각 벡터값들 간의 유사 정도를 알려주는 유사도 측정(similarity measures, 혹은 proximity, similarity coefficients라고 불림) 함수는 주요한 역할을 한다[2, 5, 11].

정보 객체간의 다양한 유사도 측정함수가 제시되어왔지만, 이들 함수간의 관계와 특징을 체계적으로 고찰한 연구는 많지 않다. 벡터 정보 객체에 대한 유사도 측정 함수 체계에 대해서도 마찬가지인데, 이 또한 유사도 측정 함수들 간의 대수적 특성이 제한되게 분석되거나 고려되지 않은 한계가 있다.

본 연구에서는 벡터 객체에서 사용되는 대표적인 유사도 측정 함수들을 고찰하되, 각각의 대수적 특성을 분석하고 서로 연결된 특성을 제시한다. 이러한 특성에 따라 유사도 측정 함수들을 분류해 본다. 이러한 과정은 표준 벡터 유사도 함수가 가져야 할 대수적 특성을 제시해준다.

본 논문의 구성은 다음과 같다. 먼저 제 2장에서는 관련 연구를, 제 3장에서는 백터 유사도 함수의 종류와 그 대수적 특성을 살펴보고, 제 4장에서는 대수적 특성을 반영한 유사도 함수의 분류에 따른 새로운 유사도 함수에 대한 제시를, 마지막으로 제 5장에서는 결론과 연구 방향을 제시한다.

2. 관련 연구

유사도 측정 함수의 종류는 적용되는 분야와 정보 객체의 표현에 따라 다양하다. 정보 객체가 집합(set)으로 표현된 경우라면 Jaccard 유사도[6]와 같은 집합 유사도를, 스트링 표현 등에서는 Levenshtein distance[9]와 같은 시퀀스 유사도를 벡터 정보 객체에서는 Salton et al.[10]의 코사인 유사도(cosine similarity) 등이 그 대표적 유사도 측정 함수의 예시라 볼 수 있다.

다양한 기존의 유사도 측정 함수들에 대한 비교 분석이 시도되어 왔고, 대부분은 특정 연산자 형식에 국한하여 유사도 함수를 고찰하고 있다[5, 1, 8, 4]. 벡터 정보 객체에 대한 유사도 함수의 비교 연구에 관한 분석적 연

구는[3, 8, 7] 정도가 대표적이라 볼 수 있다.

Cha[3]에서는 데이터의 벡터 값을 확률분포 함수로서 볼 수 있다는 점에 기반하여, 확률 통계 분야에서 폭넓게 사용되는 각종 거리(distance) 함수와 유사도 측정 함수를 분류하고 있으나, 본 논문에서 살피는 유사도 측정 함수의 대수적 특성은 분류의 요소로서 포함하고 있지 않다. Lesot et al.[8]에서는 벡터 객체에서의 유사도 측정 함수들을 거리 측정 함수와 그 구성 요소에 따라 분류하지만 그 구성요소를 프러덕트(product)에 한정하고 기타 요소들은 제외하고 있다.

Lee[7]에서는 벡터 유사도 측정 함수 체계가 포괄적으로 설명되어있고, 코사인 유사도와 같은 특정한 대수적 특성을 가진 유사도 함수를 유사도 조인(similarity join) 연산에 사용할 경우, 어떻게 최적화 시킬 수 있는지를 보여준다. 본 논문에서의 대수적 고찰은 이에 기반하는데, 오버랩(overlap) 등의 대수적 성질에 한정하고, 정보 벡터가 오피니언 마이닝과 같은 전자상거래 정보 객체에 적용될 수 있음을 보여주는 등의 차이가 있다.

3. 벡터 유사도 측정 함수

벡터는 정보 객체를 표현할 때 해당 객체의 특질(feature)에 대응하는 각 차원에 0이 아닌 값을 통해 표현한다. 웹 문서를 TF-IDF 기법을 이용하여 주요어로 이루어진 벡터 공간에 표현하는 것은 그 대표적인 예이다.

이때 특질의 객체에서의 가중치 혹은 각 특질의 객체에서의 확률 분포 등, 특질에 대응되는 각 차원의 값은 양의 값을 가지는 경우가 일반적이며, 본 논문에서도 벡터의 각 차원의 값이 0 혹은 양수인 양벡터 공간에서의 벡터 유사도 함수를 다룬다.

V를 m차원의 양벡터 공간($\mathbb{R}_0^{+m}$)이라 하면, 벡터 유사도 측정 함수는 다음과 같이 정의된다.

[정의 1] *벡터 유사도 측정 함수(VSM, vector similarity measure) S는 다음의 세 가지 조건을 만족하는 $V\times V\rightarrow\mathbb{R}_0^+$에서 정의되는 함수이다.*

- 자기 유사성(self-similarity) :
  $\forall x,\ y\in V,\ S(x,\ x)=S(y,\ y)$
- 최대성(maximality) :
  $\forall x,\ y\in V,\ S(x,\ x)\geq S(x,\ y)$
- 대칭성(symmetry) :
  $\forall x,\ y\in V,\ S(x,\ y)=S(y,\ x)$

위의 세 가지 특성은 유사도 측정 함수를 정의하는데 있어서 일반적으로 도입되는 특성이다. 최대성과 대칭성의 경우엔 몇몇 특수한 함수의 경우에는 예외적으로 만족하지 않지만, 최근에 사용되는 많은 유사도 측정 함수는 위의 세 가지 특성을 만족하고, 본 논문에서 다룰 벡터 유사도 측정 함수는 [정의 1]을 따르는 함수로 한정한다.

3.2 벡터 유사도 측정 함수의 분류

Cha[3]에서는 최근까지 주로 사용되는 많은 VSM(vector similarity measure)에 대해 소개하고 있는데, 본 논문에서는 여기에서 세

과(family)로 분류된 체계를 기반으로 VSM의 대수적 특성에 대해 살펴보기로 한다.

<Table 1>은 세 과에 속한 VSM에 대한 정의를 보여준다. x_i와 y_i는 두 벡터 x와 y의 i번째 차원에서의 벡터 값을 의미한다.

〈Table 1〉 Three Families of Vector Similarity Measures

Intersection family	
Intersection	$I(x, y) = \sum_{i=1}^{m} \min(x_i, y_i)$
Czekanowski	$S_{Cze}(x, y) = \frac{2\sum_{i=1}^{m} \min(x_i, y_i)}{\sum_{i=1}^{m} x_i + y_i}$
Motyka	$S_{Mot}(x, y) = \frac{\sum_{i=1}^{m} \min(x_i, y_i)}{\sum_{i=1}^{m} x_i + y_i}$
Kulczynski	$S_{Kul}(x, y) = \frac{\sum_{i=1}^{m} \min(x_i, y_i)}{\sum_{i=1}^{m} \lvert x_i - y_i \rvert}$
Ruzika	$S_{Ruz}(x, y) = \frac{\sum_{i=1}^{m} \min(x_i, y_i)}{\sum_{i=1}^{m} \max(x_i, y_i)}$
Dot product family	
Dot product	$D(x, y) = \sum_{i=1}^{m} x_i y_i$
Harmonic mean	$H(x, y) = \sum_{i=1}^{m} \frac{2x_i y_i}{x_i + y_i}$
Tanimoto	$S_{Tan}(x, y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} y_i^2 - \sum_{i=1}^{m} x_i y_i}$
Dice	$S_{Dice}(x, y) = \frac{2\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} y_i^2}$
Cosine	$S_{Cos}(x, y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} \sqrt{\sum_{i=1}^{m} y_i^2}}$
Fidelity family	
Fidelity	$F(x, y) = \sum_{i=1}^{m} \sqrt{x_i y_i}$

Cha[3]에서는 벡터에서의 각 차원에 대응되는 값이 확률 분포 함수(Probability Density Function, PDF)에 의해 부여되는데, 벡터의 각 차원의 모든 값의 합, 즉 L_1-norm이 1이다. 이 같이 제약된 벡터 공간 내에서는 <Table 1>에 제시된 함수는 모두 정의 1의 세 특성을 만족한다. 그러나, 본 논문에서는 L_1-norm이 1이 아닌 좀 더 일반적인 벡터, 즉 양벡터 공간상의 모든 벡터에 대해서 다룬다.

<Table 1>에 제시된 Intersection, Dot Product, Harmonic mean과 Fidelity는 [정의 1]의 자기 유사성과 대칭성은 만족하지만 최대성을 만족하지 못한다. 따라서 앞서 언급한 네 함수는 VSM으로 다루지 않고, 대신에 정의 1을 만족하는 VSM을 분류하고 새로운 VSM을 정의하는 데에 이용한다. 이 네 가지 함수는 각각 $I(x, y)$, $D(x, y)$, $H(x, y)$와 $F(x, y)$로 표현하며, 나머지 VSM은 $S_{\text{이름}}(x, y)$의 형식으로 표현한다.

3.3 벡터 오버랩에 의한 분류

<Table 1>에서 동일한 과에 속하는 유사도 함수들은 두 벡터의 동일한 차원에서의 값을 하나의 실수 값으로 결합(combine)하기 위해서 유사한 연산(operation)을 수행한다. 예를 들어, Intersection과에 속한 함수들은 두 벡터 값 중 최소값, 즉, $\min(x_i, y_i)$을 사용하고, Dot product과에 속한 함수들은 두 벡터 값의 곱, $x_i y_i$를 사용하며, Fidelity과에 속한 함수는 두 벡터 값의 기하 평균, $\sqrt{x_i y_i}$를 사용한다.

벡터 공간 상에서는 각 차원이 독립이라고 가정되고, 차원 별로 결합된 값들은 합쳐져서 하나의 스칼라(scalar) 값을 만든다. 이로부터, 두 집합의 오버랩과 유사하게 벡터 오버랩 함수를 다음과 같이 정의할 수 있다.

[정의 2] *벡터 오버랩 함수(VOM)* O*는* $V \times V \rightarrow \mathbb{R}_0^+$*에서* $O(x,y) = \sum_{i=1}^{m} \phi(x_i, y_i)$

인 함수이다. 여기서, $x, y \in V$이며 ϕ는 결합 연산자(combining operator)로 다음과 같은 네 가지 특성을 만족하는 $\mathbb{R}_0^+ \times \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$에서 정의된 함수이다.

- 0 특성(zero-property) :
  $\forall a \in \mathbb{R}_0^+, \phi(a, 0) = 0$
- 단순증가성(monotonicity) :
  $\forall a, b, c \in \mathbb{R}_0^+, b \leq c \Rightarrow \phi(a, b) \leq \phi(a, c)$
- 대칭성(symmetry) :
  $\forall x, y \in V, S(x, y) = S(y, x)$
- 동일성(identity) :
  $\forall a \in R_0^+, \phi(a, a) = a^p, where p \in \{1, 2\}$

<Table 1>에 제시된 Intersection, Dot product, Fidelity와 Harmonic mean은 네 특성을 모두 만족하는 VOM이다.

벡터 x에 대해서 그 자신과의 오버랩, 즉 $O(x, x)$를 자기 오버랩(self-overlap)이라 하고 간단히 $O(x)$로 표현하는데, 결합 연산자의 동일성(identity)으로 자기 오버랩은 벡터 놈(norm)과 유사하다. 구체적으로, 자기-Intersection, 자기-Fidelity와 자기-Harmonic mean은 벡터의 L_1-norm과 일치한다. 즉, $I(x) = F(x) = H(x) = \sum_{i=1}^{m} x_i = \| x \|_1$. 반면에 자기-Dot product는 L_2-norm의 제곱과 동일하다. 즉, $D(x) = \sum_{i=1}^{m} x_i^2 = \| x \|^2$. 이 같은 관점으로 본다면, Dot product과에 속한 Harmonic mean은 각 차원에서의 값을 결합하기 위해서 $\frac{2x_iy_i}{x_i+y_i}$를 사용하는 별도의 과로 분류되는 것이 자연스러우며, 따라서, 전자를 L_1-오버랩 그룹으로 후자를 L_2-오버랩 그룹으로 구분한다.

이렇게 하면 <Table 1>에 제시된 모든 VSM은 VOM을 이용해서 다시 표현할 수 있는데, <Table 2>에는 오버랩의 특성을 기반으로 네 개의 과(family)와 두 개의 그룹(group)으로 구분되고, VOM을 이용해서 다시 표현된 VSM이 보여주고 있다.

〈Table 2〉 Classification of Vector Similarity Measures by Overlapping Measures

L_p	Family	
L_1	Intersection family	
	Intersection	$I(x, y) = \sum_{i=1}^{m} \min(x_i, y_i)$
	Czekanowski	$S_{Cze}(x, y) = \frac{2\mathrm{I}(x_i, y_i)}{\| x \|_1 + \| y \|_1}$
	Motyka	$S_{Mot}(x, y) = \frac{\mathrm{I}(x, y)}{\| x \|_1 + \| y \|_1}$
	Kulczynski	$S_{kul}(x, y) = \frac{\mathrm{I}(x, y)}{\| x \|_1 + \| y \|_1 - 2\mathrm{I}(x, y)}$
	Ruzika	$S_{Ruz}(x, y) = \frac{\mathrm{I}(x, y)}{\| x \|_1 + \| y \|_1 - \mathrm{I}(x, y)}$
	Harmonic mean family	
	Harmonic mean	$H(x, y) = \sum_{i=1}^{m} \frac{2x_iy_i}{x_i+y_i}$
	Fidelity family	
	Fidelity	$F(x, y) = \sum_{i=1}^{m} \sqrt{x_iy_i}$
L_2	Dot product family	
	Dot product	$D(x, y) = \sum_{i=1}^{m} x_iy_i$
	Tanimoto	$S_{tan}(x, y) = \frac{D(x, y)}{\| x \|^2 + \| y \|^2 - D(x, y)}$
	Dice	$S_{Dice}(x, y) = \frac{2D(x, y)}{\| x \|^2 + \| y \|^2}$
	Cosine	$S_{Cos}(x, y) = \frac{D(x, y)}{\| x \| \| y \|}$

3.4 나눔자에 의한 분류

몇몇 VSM은 동일한 과의 다른 VSM에 의해서 계산될 수 있다. Intersection과의 $S_{Mot}(x, y)$, $S_{Kul(x,y)}$와 $S_{Ruz}(x, y)$는 식 (1)~식 (3)에서 보이는 바와 같이 $S_{Cze}(x, y)$로 계산할 수 있고, Dot product과의 $S_{Tani}(x, y)$는 식 (4)에서 보이는 바와 같이 $S_{Dice}(x, y)$로 계산할 수 있다.

$$S_{Mot}(x, y) = \frac{1}{2} S_{Cze}(x, y) \quad (1)$$

$$S_{Kul}(x, y) = \frac{S_{Cze}(x, y)}{2 - 2S_{Cze}(x, y)} \quad (2)$$

$$S_{Ruz}(x, y) = \frac{S_{Cze}(x, y)}{2 - S_{Cze}(x, y)} \quad (3)$$

$$S_{Tani}(x, y) = \frac{S_{Dice}(x, y)}{2 - S_{Dice}(x, y)} \quad (4)$$

다른 VSM에 의해서 정의될 수 있는 VSM들 이외의 함수들에는 두 벡터의 오버랩을 각 벡터의 자기 오버랩을 이용해서 표준화(normalize)로 나타낼 수 있데, $S_{Cze}(x, y)$, $S_{Dice}(x, y)$, $S_{Cos}(x, y)$는 식 (5)~식 (7)에서 보이는 바와 같이 쓰일 수 있다. 여기서, $\boldsymbol{A}(a, b) = \frac{(a+b)}{2}$ 이고 $\boldsymbol{G}(a, b) = \sqrt{ab}$ 이다.

$$S_{Cze}(x, y) = \frac{I(x, y)}{\boldsymbol{A}(I(x), I(y))} \quad (5)$$

$$S_{Dice}(x, y) = \frac{D(x, y)}{\boldsymbol{A}(D(x), D(y))} \quad (6)$$

$$S_{Cos}(x, y) = \frac{D(x, y)}{\boldsymbol{G}(D(x), D(y))} \quad (7)$$

즉, VSM은 두 벡터의 오버랩을 어떻게 표준화 하는가에 의해서도 분류될 수 있다. 구체적으로 Dice 유사도와 Czekanowski 유사도는 산술 평균(arithmetic mean) 나눔자 그룹, Cosine 유사도는 기하 평균(geometric mean) 나눔자 그룹에 속한다. 이 같은 구분 방법은 오버랩이 각 차원 값의 결합 방법에 따라 분류되는 것과 달리 VSM의 구조적 특성에 의한 분류이다.

4. 새로운 벡터 유사도 측정 함수

Czekanowski, Ruzika, Dice, Tanimoto와 Cosine 등의 유사도는 두 벡터가 더 유사할수록 1에 가까운 수를, 두 벡터가 다를수록 0에 가까운 수를 반환한다. 이들 VSM을 표준 벡터 유사도 측정 함수(SVSM, standard vector similarity measure)이라 하면, 이들 외에도 VSM의 대수적 특성을 이용하여 추가로 SVSM을 제시할 수 있다.

4.1 벡터 오버랩 함수

T^p를 m차원 벡터 x를, $T^p(x) = \langle x_1^p, \cdots, x_i^p, \cdots, x_m^p \rangle$로 변환하는 $V \mapsto V$인 변환이라 하자. 그러면, $D(x, y) = F(T^2(x), T^2(y))$, 즉, $\sum_{i=1}^{m} x_i y_i = \sum_{i=1}^{m} \sqrt{x_i^2} y_i^2$. 마찬가지로, Harmonic mean과 Intersection에 대응하는 벡터 오버랩 함수 $H(T^2(x), T^2(y))$와 $I(T^2(x), T^2(y))$를 정의할 수 있다. 단순히 VOM을 O_p와 같이 표현하자. 이때, $O \in \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{I}$이고 $p \in 1, 2$. 예로, $\boldsymbol{G_1}$은 Fidelity를 나타내는데, 각 벡터 값을 결합하기 위해 기하 평균을 사용하고, $\boldsymbol{G_2}$는 Dot

product를 나타내며 각 벡터 값을 제곱하여 기하 평균을 이용하여 결합한다. <Table 3>은 이 같은 방법으로 정의할 수 있는 벡터 오버랩 함수를 보여준다.

〈Table 3〉 Vector Overlap Measures

	L_1-overlaps		L_2-overlaps	
Geometric mean	G_1	$\Sigma_{i=1}^{m} \sqrt{x_i y_i}$	G_2	$\Sigma_{i=1}^{m} \sqrt{x_i^2 y_i^2}$
Harmonic mean	H_1	$\Sigma_{i=1}^{m} \frac{2x_i y_i}{x_i + y_i}$	H_2	$\Sigma_{i=1}^{m} \frac{2x_i^2 y_i^2}{x_i^2 + y_i^2}$
Intersection(Min)	I_1	$\Sigma_{i=1}^{m} \min(x_i, y_i)$	I_2	$\Sigma_{i=1}^{m} \min(x_i^2, y_i^2)$

이후 결합 연산자를 나타내기 위해서도 동일한 표현 방법을 사용한다. 예로, 두 벡터 x와 y에 대해 $\boldsymbol{H}_1(x, y) = \Sigma_{i=1}^{m} \frac{2x_i y_i}{x_i + y_i}$ $\boldsymbol{H}_1(x, y)$는 벡터 오버랩 함수, 즉, 이고, 두 실수 a와 b에서 $\boldsymbol{H}_1(a, b)$는 결합 연산자, 즉, $\boldsymbol{H}_1(a, b) = \frac{2ab}{a+b}$ 이다.

4.2 나눔자

N을 벡터 유사도 측정 함수에서의 나눔자로 표현하자. 그러면, VSM은 $\frac{O_p(x, y)}{N(O_p(x), O_p(y))}$과 같이 단순하게 표현될 수 있고, [정의 1]의 자기 유사성에 의해 $N(O_p(x), O_p(x)) = kO_p(x)$임을 알 수 있다($k \in \mathbb{R}_0^+$). $k=1$이면 N에 의해 정의된 VSM은 SVSM이다. 제 3.4절에서 제시된 산술 평균(A)과 기하 평균(G) 외에도 <Table 4>에 보여지는 바와 같이 조화 평균(H)과 Intersection(I) 또한 나눔자로 사용될 수 있다.

〈Table 4〉 Denominators

Arithmetic mean	A	$\frac{O_p(x) + O_p(y)}{2}$
Geometric mean	G	$\sqrt{(O_p(x) O_p(y)}$
Harmonic mean	H	$\frac{2O_p(x) O_p(y)}{O_p(x) + O_p(y)}$
Intersection(Min)	I	$\min(O_p(x), O_p(y))$

4.3 오버랩 함수와 나눔자의 조합

앞의 두절에서 정리된 VOM과 나눔자를 이용하면 총 24개의 조합을 얻을 수 있다. 편리성을 위해 VOM O_p와 나눔자 N의 조합에 의한 VSM을 S_{O_pN}과 같이 나타낸다($O \in \{\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{I}\}, p \in \{1, 2\}, N \in \{\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{I}\}$). 예를 들어, S_{I_1A}, S_{G_2I}와 S_{G_2G}는 각각 Czekanowski, Dice와 Cosine 유사도를 나타낸다.

24개의 모든 조합은 자기 유사성($\forall x \in V, S_{O_pN}(x, x) = 1$)과 대칭성을 만족한다. 그러나 몇몇 조합은 최대성(maximality)을 만족하지 않는다. 예로 $\boldsymbol{G}_1$과 $\boldsymbol{I}$의 조합을 보자. 만약 두 벡터가 x = <0.2, 0.3>와 y = <0.5, 0.4>라면, $S_{G_1I}(x, y) = \frac{G_1(x, y)}{I(\| x \|_1, \| y \|_1)} = \frac{\sqrt{0.2 \times 0.5} + \sqrt{0.3 \times 0.4}}{\min(0.2+0.3, 0.5+0.4)} \simeq 1.3$. 따라서, S_{G_1I}는 최대성을 만족하지 않는다.

자기 유사성($\forall x \in V, S_{O_pN}(x, x) = 1$)과 최대성($\forall x, y \in V, S_{O_pN})(x, y) \leq S_{O_pN}(x, x)$)으로부터 $\forall x, y \in V, O_p(x, y) \leq N(O_p(x), O_p(y))$임을 알 수 있다. 이는 벡터 유사도 측정 함수가 최대성을 만족하기 위해서는, 두 실수가 주어졌을 때 나눔자가 결합 연산자보다 크거나 같

은 값을 반환해야 한다는 것을 의미한다. VOM의 특성과 나눔자의 특성을 이용해서 SVSM을 만드는 조합을 찾을 수 있다.

먼저 두 벡터의 오버랩의 상한을 두 벡터의 자기오버랩을 이용해서 [정리 1]과 같이 제한할 수 있다. 지면 관계 상 증명은 생략하겠다. 이에 대한 상세한 내용은 Lee[7]을 참고하기 바란다.

[정리 1] *각 차원에서의 값을 결합하기 위해서 연산자 ϕ를 사용하는 벡터 오버랩 함수 O_p에 대해서 항상 다음이 성립한다 :*

$$\forall x, y \in V,\ O_p(x, y) \le \phi(\sqrt[p]{O_p(x)},\ \sqrt[p]{O_p(y)}) \tag{8}$$

이로부터 벡터 오버랩의 상한은 나눔자에 대응시킬 수 있다. 예로, $G_2(x, y) \le G_2(\sqrt{G_2(x)}, \sqrt{(G_2(y))} = G(G_2(x),\ G_2(y))$. 따라서 벡터 오버랩 함수 $\boldsymbol{G_2}$와 나눔자 $\boldsymbol{G}$로 정의되는 함수는 최대성을 만족한다.

이 외에도, [보조 정리 1]에 의해 벡터 오버랩 함수 $\boldsymbol{G_2}$와 나눔자 $\boldsymbol{A}$의 조합도 최대성을 만족한다.

[보조 정리 1]

$$\forall a, b \in \mathbb{R}_0^+,\ I(a, b) \le H(a, b) \le G(a, b) \le A(a, b) \tag{9}$$

이 같은 방법으로 총 18개의 SVSM을 만들어내는 조합을 찾아낼 수 있다. <Table 5>

〈Table 5〉 Standard Vector Similarity Measures

Overlap \ Denominator	L_1-overlap group			L_2-overlap group
	Fidelity family	Harmonic mean family	Intersection family	Dot product family
	Fidelity	Harmonic mean	Intersection	Dot product
	$G_1(x,\ y) = \sum_{i=1}^{m}\sqrt{x_i y_i}$	$H_1(x,\ y) = \sum_{i=1}^{m}\frac{2x_i y_i}{x_i + y_i}$	$I_1(x,\ y) = \sum_{i=1}^{m}\min(x_i,\ y_i)$	$G_2(x,\ y) = \sum_{i=1}^{m} x_i y_i$
	$G_1(x) = \Vert x \Vert_1$	$H_1(x) = \Vert x \Vert_1$	$I_1(x) = \Vert x \Vert_1$	$G_2(x) = \Vert x \Vert^2$
Arithmetic mean	S_{G_1A}	S_{H_1A}	S_{I_1A}, Czekanowski	S_{G_2A}, Dice
$\frac{O(x,\ y)}{\frac{O(x)+O(y)}{2}}$	$\frac{G_1(x,\ y)}{\frac{\Vert x \Vert_1 + \Vert y \Vert_1}{2}}$	$\frac{H_1(x,\ y)}{\frac{\Vert x \Vert_1 + \Vert y \Vert_1}{2}}$	$\frac{I_1(x,\ y)}{\frac{\Vert x \Vert_1 + \Vert y \Vert_1}{2}}$	$\frac{G_2(x,\ y)}{\frac{\Vert x \Vert^2 + \Vert y \Vert^2}{2}}$
Geometric mean	S_{G_1G}, General Fidelity	S_{H_1G}	S_{I_1G}	S_{G_2G}, Cosine
$\frac{O(x,\ y)}{\sqrt{O(x)O(y)}}$	$\frac{G_1(x,\ y)}{\sqrt{\Vert x \Vert_1 \Vert y \Vert_1}}$	$\frac{H_1(x,\ y)}{\sqrt{\Vert x \Vert_1 \Vert y \Vert_1}}$	$\frac{I_1(x,\ y)}{\sqrt{\Vert x \Vert_1 \Vert y \Vert_1}}$	$\frac{G_2(x,\ y)}{\Vert x \Vert \Vert y \Vert}$
Harmonic mean		S_{H_1H}	S_{I_1H}	
$\frac{O(x,\ y)}{\frac{2O(x)O(y)}{O(x)+O(y)}}$		$\frac{H_1(x,\ y)}{\frac{2\Vert x \Vert_1 \Vert y \Vert_1}{\Vert x \Vert_1 + \Vert y \Vert_1}}$	$\frac{I_1(x,\ y)}{\frac{2\Vert x \Vert_1 \Vert y \Vert_1}{\Vert x \Vert_1 + \Vert y \Vert_1}}$	
Intersection			S_{I_1I}	
$\frac{O(x,\ y)}{\min(O(x),\ O(y))}$			$\frac{I_1(x,\ y)}{\min(\Vert x \Vert_1,\ \Vert y \Vert_1)}$	

에는 H_2와 I_2 벡터 오버랩 함수에 의한 조합을 제외한 총 11개의 SVSM이 정리되어 있다. 새로운 SVSM S_{G_1G}는 L_1-norm이 1로 표준화된 벡터들에 대해서는 Fidelity와 동일한 값을 만들기 때문에 Cha[3]에서의 Fidelity와 상응하며, S_{G_1G}를 일반화된 Fidelity(General Fidelity)라 칭한다.

5. 맺음말

유사도 측정 함수는 정보 객체간 중복 및 연관성 파악에 사용되며 전자상거래 분야 등 광범위한 분야에서 사용된다. 본 논문에서는 정보 객체가 벡터로 표현될 때 사용되는 유사도 측정 함수를 그 대수적 특성에 따라 살펴보고 분류해 보았다. 주로 고려된 특성은 함수의 구성 요소이고 특히 벡터오버랩을 기준으로 유사도 함수 체계를 살펴보았다.

유사도 함수의 대수적 특성을 분석해보고 그 특징을 발견하려는 연구는 단순히 기존에 발표된 함수 성질 파악에 국한하지 않고, 새로운 유사도 함수의 제안에 사용될 수 있다. 이를 위해서는 제안된 새로운 유사도 함수가 특정 도메인에서 어떠한 효과를 보이는지에 대한 실험 및 그 성능 평가가 중요함은 물론이다.

빅데이타가 사회적, 기술적으로 그 중요도가 늘어가고 있는 시점에서, 광대한 사용자의 리뷰 문서에서 필요한 의견을 추출 분석하는 오피니언 마이닝 분야 등에서 효과적인 유사도 함수와 연산으로의 효과적 적용 등은 앞으로 연구가 더욱 필요하다.

References

[1] Batagelj, V. and Bren, M., “Comparing resemblance measures,” Journal of Classification, Vol. 12, 1995.

[2] Bouchon-Meunier, B., Rifqi, M., and Bothorel, S., “Towards general measures of comparison of objects,” Fuzzy Sets Systems, Vol. 84, 1996.

[3] Cha, S.-H., “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions,” INTERNATIONAL JOURNAL of MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, Vol. 1, 2007.

[4] Choi, S.-S., Cha, S.-H., and Tappert, C. C., “A Survey of Binary Similarity and Distance Measures,” Journal of Systemics, Cybernetics and Informatics, Vol. 8, 2010.

[5] Deza, M.-M. and Deza, E., Dictionary of Distances, Elsevier Science, 2006.

[6] Jaccard, P., “Étude comparative de la distribution florale dans une portion des Alpes et des Jura,” Bulletin del la Société Vaudoise des Sciences Naturelles, 1901.

[7] Lee, D., An Efficient Filtering Framework for Vector Similarity Joins, Ph.D. Thesis, Seoul National University, 2011.

[8] Lesot, M.-J., Rifqi, M., and Benhadda, H., “Similarity measures for binary and numerical data : a survey,” International Journal of Knowledge Engineering and

Soft Data Paradigms, Vol. 1, 2009.

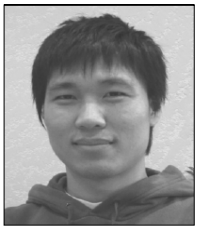
[9] Levenshtein, V., “Binary codes capable of correcting deletions, insertions and reversals,” Soviet Physics Doklady, Vol. 10, 1966.

[10] Salton, G., Wong, A., and Yang, C. S., “A vector space model for automatic indexing,” Communications of the ACM, Vol. 18, 1975.

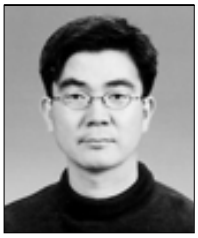
[11] Santini, S. and Jain, R., “Similarity Measures,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, 1999.

[12] Yeon, J., Lee, D., Shim, J., and Lee, S.-G., “Product Review Data and Sentiment Analytical Processing Modeling,” The Journal of Society for e-Business Studies, Vol. 16, 2011.

저 자 소 개



이동주 (E-mail : therocks@europa.snu.ac.kr)
2003년 서울대학교 응용생물화학부 졸업 (학사)
2011년 서울대학교 컴퓨터공학부 대학원 (박사)
2011년~현재 삼성전자 DMC연구소 책임연구원
관심분야 데이터베이스, 자연어 처리, 상황인지 개인화



심준호 (E-mail : jshim@sookmyung.ac.kr)
1990년 서울대학교 계산통계학과 졸업 (학사)
1994년 서울대학교 계산통계학과 전산과학전공 (석사)
1998년 Northwestern University, Electrical and Computer Engineering (박사)
2001년~현재 숙명여자대학교 컴퓨터과학부 교수
관심분야 데이타베이스, 전자상거래, 데이타웨어하우스, 빅데이타