An Axiomatic Account of Similarity

Enrique Amigó, Julio Gonzalo, Fernando Giner and Felisa Verdejo
enrique,julio,felisa@lsi.uned.es
National Distance University
C Juan del Rosal 16
Madrid 28040

ABSTRACT

Although computing similarity is one of the fundamental challenges of Information Access tasks, the notion of similarity is not yet completely understood from a formal, axiomatic perspective. In this paper we show how axiomatic explanations of similarity from other fields (i.e. Tversky's axioms from the point of view of cognitive sciences, and metric spaces from the point of view of algebra) do not completely fit the problem of similarity in Information Access, and we propose a new set of axioms which can be synthesized into a single *Similarity Information Monotonicity* axiom (SIM). Directly grounded on SIM, we then introduce a new similarity model, the *Information Contrast Model*, which generalizes both Tversky's linear contrast model and Pointwise Mutual Information, and, unlike previous similarity models, satisfies the SIM axiom for a certain range of values of its parameters.

CCS CONCEPTS

•Information systems → Document representation; Similarity measures;

ACM Reference format:
Enrique Amigó, Julio Gonzalo, Fernando Giner and Felisa Verdejo. 2017. An Axiomatic Account of Similarity. In *Proceedings of SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks, Tokyo Japan, August 2017 (ATIR),* 10 pages.
DOI:

1 INTRODUCTION

Information retrieval systems are (at least partially) based on computing the similarity between query and documents. Summarization, Clustering and many other text processing applications require computing the similarity between texts. Evaluation measures for text generation tasks (such as summarization or machine translation), computing textual similarity is the key to compare the output of systems with the models produced by humans. And, beyond textual similarity, applications such as collaborative recommendation are based on estimating the similarity between users (based on their preferences and behaviour) and between products (based on also on user preferences). In summary, computing similarity is a core problem which pervades, either implicitly or explicitly, many Information Access tasks.

In general, computing similarity deals with two problems: (i) how best to represent (and possibly enrich representations of) objects; and (ii) how best to compare object representations. In this paper we will focus on the second step: once two objects are represented (as sets of suitable features), how should we compute similarity between that representations? For the sake of clarity, we will illustrate our analysis in terms of texts represented as bags of words, but our formal study abstracts from which kinds of features are used to represent objects (words, n-grams, concepts, user preferences, syntactic/semantic relationships...) and concentrates on the problem of comparing representations to infer similarity. Our main goal is to deepen into the notion of similarity by providing a suitable axiomatic characterization.

The closest references to model similarity in Information Access come from Algebra (the notion of distance in metric spaces, which play a role in many text processing models), from Information Theory based models and from Cognitive Science (most notably Tversky's work on conceptual similarity). We will see, however, that axiomatics from these fields are not entirely suitable to explain the notion of similarity in the context of Information Access in general; indeed, counterexamples can be found for many of the similarity axioms proposed in the past.

Our first contribution in this paper is to **define a new axiomatic account of similarity based on concepts from Information Theory**. We first postulate four intuitive axioms; *identity, identity specificity, unexpectedness, dependency and asymmetricity*. All of them can be derived from a single *Similarity Information Monotonicity (SIM)* axiom. A literature review shows that none of the existing similarity models is able to satisfy our basic axioms, although different techniques at the representation level may mitigate their potential problems.

Our second contribution is to **propose a new model to compute textual similarity**, the *Information Contrast Model* (ICM), which derives directly from the SIM axiom. ICM computes similarity between two objects as a linear combination of the individual information quantity of each object and the information quantity of its (multi set) union. The model generalizes multiple approaches such as PointWise Mutual Information, Tversky's linear contrast model, language models or conditional probability. And, most importantly, ICM satisfies our formal axioms for a specific range of values of its parameters.

ATIR, Tokyo Japan

DOI:

2 PREVIOUS AXIOMATIC FRAMEWORKS

2.1 Metric Spaces

The most traditional axiomatic framework comes from the concept of metric space [15, 32]. In Psychology, the assumption that similarity can be expressed as distances in a metric space is known as the *Generalization Law* [31]. The first axiom is *maximality*, which states that every pair of identical objects achieve a maximal and constant similarity:

$$Sim(\mathcal{X},\mathcal{X}) = Sim(\mathcal{Y},\mathcal{Y}) \geq Sim(\mathcal{X},\mathcal{Y})$$

However, maximality has already been objected in the context of cognitive sciences [14]. Based on several experiments – e.g. cognition of Morse code [30] and cognition of rectangles varying in size and reflectance [4], – many researchers claimed that the axiom of maximality does not correspond with human intuitions. In particular, Tversky's experiments showed that maximality (or minimality in distance) does not hold if a larger stimulus with more signs is compared to a smaller one with less signs: if a stimulus shows more details, its level of perceived self-similarity increases [34]. In the context of Information Retrieval (IR) this phenomenon is correlated with query specificity. For instance, we can not ensure that a document containing only the words *"Good news"* satisfies the user query *"Good news"*. But typing the full content of an article as a query ensures the relevance of the article as retrieved document. In both cases, we are talking about self-similarity, but in the second case the object contains more information.

The second axiom is *triangular inequality*:

$$Sim(\mathcal{X},\mathcal{Y}) \leq Sim(\mathcal{X},\mathcal{Z}) + Sim(\mathcal{Z},\mathcal{Y})$$

which has also been refuted in several cognitive experiments [29, 30]. Other studies also found evidence against the third axiom, *symmetricity* ($Sim(\mathcal{X},\mathcal{Y}) = Sim(\mathcal{Y},\mathcal{X})$) [3, 35]. From a cognitive point of view, the reason is that, according to human perceptions, specific concepts tend to be closer to generic concepts than viceversa. For instance, Tversky found that subjects perceived the concept *"North Korea"* as being closer to *"China"* than vice versa, because China has more salient distinctive features than North Korea. This is also valid for the textual similarity context.

2.2 Axiomatics of Tversky and Gati

Tversky and Gati [35] tried to state axiomatics for similarity from an ordinal perspective, defining a *monotone proximity structure* which is based on three properties. The first one is *dominance*, which states that replacing a different feature by a common feature increases similarity:

$$Sim(\mathcal{X}_1\mathcal{Y}_1,\mathcal{X}_2\mathcal{Y}_2) < min\{Sim(\mathcal{X}_1\mathcal{Y}_1,\mathcal{X}_1\mathcal{Y}_2), Sim(\mathcal{X}_1\mathcal{Y}_1,\mathcal{X}_2\mathcal{Y}_1)\}$$

Exemplified with words as features, this implies that Sim("brown monkey","red cross") is lower than Sim("**brown** monkey","**brown** cross") because the second case texts share one feature.

However, this axiom is grounded on the idea of independence across dimensions, but words – and other features – do not co-occur randomly. For instance:

Example 2.1.

$$Sim(\text{"}Disney\ mouse\text{"},\text{"}game\ Mickey\text{"}) > Sim(\text{"}Disney\ \textbf{mouse}\text{"},\text{"}game\ \textbf{mouse}\text{"})$$

Even if they do not share any word, *"Mouse Disney"* can be closer to *"game Mickey"* than *"Disney mouse"* to *"game mouse"*, contradicting the dominance axiom. Notice that Mickey is commonly associated with the Disney character, while *"game mouse"* can be associated with other contexts, for instance computer mouses and games.

The second axiom is *consistency*, which states that the ordinal relation between similarities along one dimension is independent of the other dimension.

$$Sim(\mathcal{X}_1\mathcal{Y}_1,\mathcal{X}_2\mathcal{Y}_1) < Sim(\mathcal{X}_3\mathcal{Y}_1,\mathcal{X}_4\mathcal{Y}_1) \Leftrightarrow Sim(\mathcal{X}_1\mathcal{Y}_2,\mathcal{X}_2\mathcal{Y}_2) < Sim(\mathcal{X}_3\mathcal{Y}_2,\mathcal{X}_4\mathcal{Y}_2)$$

Again, this axiom is grounded on the assumption that features are mutually independent. We can found also counter samples for this in the context of textual similarity. For instance, the word *"mouse"* is closer to *"Mickey"* than to *"hardware"* in the context of *"Disney"*, but not in the context of computers and external devices (*"Wireless"*).

Example 2.2.

$$Sim(\text{"Mouse }\textbf{Disney}\text{"},\text{"Mickey }\textbf{Disney}\text{"}) > Sim(\text{"Mouse }\textbf{Disney}\text{"},\text{"Hardware }\textbf{Disney}\text{"})$$
$$Sim(\text{"Mouse }\textbf{Wireless}\text{"},\text{"Mickey }\textbf{Wireless}\text{"}) < Sim(\text{"Mouse }\textbf{Wireless}\text{"},\text{"Hardware }\textbf{Wireless}\text{"})$$

The third constraint, *transitivity*, is grounded on a definition of *"betweenness"* which assumes the validity of *consistency*. Therefore, Example 2.2 also contradicts this third axiom.

2.3 Feature Contrast Model

The most popular study of Tversky [34] about similarity is the *Feature Contrast Model*. Assuming that objects can be represented as sets of features, he defined three basic axioms: *matching*, *monotonicity* and *independence*. Once more, all of them are based on the idea that features are mutually independent. *Matching* states that similarity can be computed as a function of the intersection and difference. *Monotonicity* is closely related with *Dominance*,. It states that increasing the intersection or decreasing the difference between sets, increases the similarity. But, again, we know this is not always true for texts. Because words do not occur independently from each other, adding different words to a pair of texts may increase their similarity, as in this example where *"Desktop"* and *"Computer"* bring *"Apple"* and *"Mouse"* to the context of computers.

Example 2.3.

$$Sim(\text{"Apple Desktop"},\text{"Mouse Computer"}) > Sim(\text{"Apple"},\text{"Mouse"})$$

Example 2.1 (from previous section) also contradicts monotonicity, given that similarity increases in spite of the fact that the intersection decreases and the difference increases.

The third property is (*independence*). Its formalization is less intuitive than other axioms. It states that, being the intersection

($\mathcal{X} \cap Y$) and the differences ($\mathcal{X} \setminus \mathcal{Y}$,$\mathcal{Y} \setminus \mathcal{X}$) the three components of similarity, if ($\mathcal{X}$,$\mathcal{Y}$) and ($\mathcal{X}'$,$\mathcal{Y}'$), share the same two components as ($\mathcal{W}$,$\mathcal{Z}$) and ($\mathcal{W}'$,$\mathcal{Z}'$), while ($\mathcal{X}$,$\mathcal{Y}$) and ($\mathcal{W}$,$\mathcal{Z}$) share a third component as well as ($\mathcal{X}'$,$\mathcal{Y}'$) and ($\mathcal{W}'$,$\mathcal{Z}'$), then:

$$Sim(\mathcal{X},\mathcal{Y}) > Sim(\mathcal{W},\mathcal{Z}) \leftrightarrow Sim(\mathcal{X}',\mathcal{Y}') > Sim(\mathcal{W}',\mathcal{Z}')$$

Example 2.2 also contradicts this property. Note that the first and third similarity instances share the difference sets ($\mathcal{X} \setminus \mathcal{Y}$ ="*Mouse*" and $\mathcal{Y} \setminus \mathcal{X}$ ="*Mickey*"). The second and fourth similarity instances also share the difference, and the first and second, as well as the third and fourth share the intersection component ("*Disney*" and "*Wireless*"). The independence axiom is violated because a human may understand that "*Mouse*" is closer to "*Mickey*" than to "*Hardware*" in the context of "*Disney*" but not in the context of "*Wireless*".

2.4 Axiomatics in Information Retrieval

Information retrieval is grounded on similarity principles, in the sense that the basic IR scenario can be interpreted as the problem of estimating the similarity between a certain user query and documents in a collection. Fang and Zhai presented a seminal work about the axiomatics of information retrieval. In their first proposal [8] they stated six axioms, which were eventually refined to only three general constraints in [9]: (i) Adding one query term to a document must increase the score, (ii) adding a non-query term to a document must decrease the score and (iii) the amount of increase in the score due to adding a query term to a document must decrease as we add more and more query terms.

The first and second axioms are, in fact, equivalent to Tversky's monotonicity axioms when we interpret the occurrence of words as features. The third constraint is an extension that determines the effect of new common features in similarity when they are added progressively. In any case, accepting the first axiom is necessary to assume the third one. Therefore, the counter example shown for Tversky's monotonicity axiom also hold here, and they derive from the need of assuming independence between words.

In summary, according to our analysis, existing axiomatics do not fit the concept of similarity in the context of information access, and one of the main reasons is that previous proposals tend to assume independence between features, a condition that is not always met. In the following section we propose a new set of axioms that take this problem into account.

3 PROPOSAL: AXIOMATICS FOR SIMILARITY IN INFORMATION ACCESS SCENARIOS.

3.1 Notation and Representation

3.1.1 Objects as Multisets of Features. Let us assume that an object is represented as a multi-set (a set with possibly repeated elements) $\mathcal{X}$ of observed features belonging to a feature domain Ω of features, i.e. $\mathcal{X} \equiv \{x_1, x_2.., x_n\} \in \Omega^n$. For simplicity, in all our examples features will be words; but all our reasoning is feature-agnostic and equally valid for other representation features of information pieces, such as n-grams, concepts, syntactic and semantic relationships, meta-data, user preferences in recommendation scenarios, followers in a network, etc.

3.1.2 Operators. Therefore, we can apply the multi-set union, intersection and inclusion operators over objects. Note that, according to the definition of multi-set, the union and intersection rules correspond with the maximal and minimal cardinalities: {abb}∪{cb}) = {abbc} and {abb} ∩ {cb} = {abc}. We will also use the multi-set sum operator, $\mathcal{X} + \mathcal{Y}$, using the simplified notation $\mathcal{X}\mathcal{Y}$ to denote it. Using the same sample sets, {abb} + {cb} = {abbcb}. If both multi-sets have no features in common, their sum is equivalent to their union.

3.1.3 Probabilistic Space. Let us consider a set of information object samples, $\mathcal{D}$, where each sample, $d \in \mathcal{D}$, is represented as a feature multi-set. in order to model the problem of similarity in probabilistic terms, we need to interpret these multi-sets as events. In the literature, this probabilistic space have been defined in several ways. For instance, the traditional IR models uses the space of documents as events; language models consider the potential word sequences that could be generated, and word embeddings (such as word2vec, Glove, etc.) consider contextual windows around the word to be represented as sample set.

In general, features are accumulative. For instance, observing a sequence of words $\{abc\}$ implies observing the subsequence $\{ab\}$ in terms of language models. Therefore, considering objects as multi-sets of observed features, we can generalize the likelihood of an object as the probability of observing a superset in the sample space: $(P(\mathcal{X}) = P_{d\in\mathcal{D}}(\mathcal{X} \subset d))$. Accordingly, the joint probability of two texts $\mathcal{X}$ and $\mathcal{Y}$ is the probability of finding the union of both feature multi-sets:

$$P(\mathcal{X},\mathcal{Y}) = P(\mathcal{X} \cup \mathcal{Y}) = P_{d\in\mathcal{D}}(\mathcal{X} \cup \mathcal{Y} \subset d)$$

The idea is also valid for other contexts. For instance, in collaborative recommendation a "like" from an user is a product feature. The probability associated to this feature would be the probability of a product to achieve a "like" from this user.

Note that the concept of probability, and therefore similarity, is related with the nature of the sample set. Therefore, different sample sets lead to different notions of similarity. Considering large contexts (e.g. full documents may provide evidence about topical similarity (which is the kind of similarity used in Information Retrieval), but considering small contexts may provide evidence about interchangeability (as in word embedding models). Our axioms do not prescribe what are the events (or contexts), and therefore can accommodate different notions of similarity depending on how events are defined.

3.2 Axioms

The first intuitive idea is that changing an object (by removing or adding information) decreases its similarity with the original. In an intuitive manner, *"if something changes, it is not the same anymore"*. Formally:

AXIOM 1. **Identity Axiom:** *Adding or removing features to an object decreases the similarity to the original:*

$$Sim(\mathcal{X},\mathcal{X}) > Sim(\mathcal{X},\mathcal{X}\mathcal{Y}) \text{ and } Sim(\mathcal{X}\mathcal{Y},\mathcal{X}\mathcal{Y}) > Sim(\mathcal{X}\mathcal{Y},\mathcal{X})$$

For instance, although we can not state axiomatically how close is *"Apple Mouse"* to itself, we can at least say that it is more similar to itself than to *"Apple"* or to *"Apple Mouse Desktop"*. This axiom is actually a relaxed version of maximality: we postulate that any

object is more similar to itself than to any other object, but not that its self-similarity is maximal.

The reason to avoid postulating *Maximality* is that, according to Tversky's findings, more informative texts are more self-similar. We reflect this idea with a second axiom:

AXIOM 2. **Identity Specificity Axiom:** *Adding new features to a text increases its self-similarity. Being* $Y \neq \emptyset$*:*

$$Sim(\mathcal{X}\mathcal{Y},\mathcal{X}\mathcal{Y}) > Sim(\mathcal{X},\mathcal{X})$$

For instance, in the context of Information Retrieval, being a query exactly like a document content, the more both contain information, the more the relevance is ensured.

We now depart from Tversky's axioms. We have seen that their main drawback, in the context of textual similarity, is that Tversky does not take into account the dependencies between features across intersection and differences of objects [10]. Therefore, **instead of grounding axioms on the independence assumption, we formalize dependency with two new axioms**. The first one states that adding a new feature decreases similarity to a greater extent if it is unexpected. For instance, *"Mickey mouse"* is closer to *"Mickey"* than *"Mickey Apple"*, because *"Apple"* is less expected in the context of *"Mickey"* than *"mouse"*.

AXIOM 3. **Unexpectedness Axiom:** *An added feature changes the text more if it is less expected:*

$$If\ P(\mathcal{Y}|\mathcal{X}) < P(\mathcal{Y}'|\mathcal{X})\ then$$
$$Sim(\mathcal{X},\mathcal{X}\mathcal{Y}) < Sim(\mathcal{X},\mathcal{X}\mathcal{Y}')$$

We also want to incorporate the possibility that adding different features to different objects may bring them together (instead of necessarily making them less similar as it is postulated by Tversky). For instance, Sim(*"Apple Desktop"*, *"Mouse Computer"*) > Sim(*"Apple"*,*"Mouse"*). We postulate that this happens when the conditional probabilities of finding one text given the other (and viceversa) increase:

AXIOM 4. **Dependency Axiom:** *Adding new features in both objects increases their similarity whenever their respective conditional probabilities grow:*

$$If\ P(\mathcal{X}\mathcal{Z}|\mathcal{Y}\mathcal{Z}') > P(\mathcal{X}|\mathcal{Y})$$
$$and\ P(\mathcal{Y}\mathcal{Z}'|\mathcal{X}\mathcal{Z}) > P(\mathcal{Y}|\mathcal{X})$$
$$then\ Sim(\mathcal{X}\mathcal{Z},\mathcal{Y}\mathcal{Z}') > Sim(\mathcal{X},\mathcal{Y})$$

For instance, in the case of IR, suppose that there is no correspondence between a query and a document: different words, different domain. So, the probabilities $P(q|d)$ and $P(d|q)$ are extremely low. Then we add the name of an artist in q and his artistic name in d, then both $P(q|d)$ and $P(d|q)$ will be higher and the estimated similarity should increase.

Notice that Tversky's monotonicity axiom is not compatible with the dependency axiom given that a new feature in one text represents an increase in the difference component (see Proof 1 in the additional material).

Finally, there are multiple studies confirming that similarity is inherently asymmetric [34], and that specific objects are closer to general objects than vice versa. Assuming that text specificity grows when adding features, we formalize an *asymmetricity* axiom in the following way:

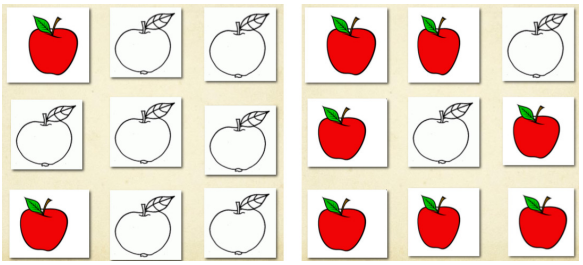

Figure 1: Red and white apples are considered the most similar object pairs by humans in the left and right side, respectively.

AXIOM 5. **Asymmetricity:** *An object is more similar to any of its parts than viceversa:*

$$Sim(\mathcal{X}\mathcal{Y},\mathcal{X}) > Sim(\mathcal{X},\mathcal{X}\mathcal{Y})$$

4 SIMILARITY INFORMATION MONOTONICITY AXIOM

We will now show that we can join previous axioms into an unique axiom that we call *Similarity Information Monotonicity* axiom (SIM).

4.1 Intuitions and Formalization

SIM is based on two main intuitions. The first one is that **the proximity of objects is related with their Information Quantity** Let us consider Figure 1. When asking people for the most similar pair of apples, most of them answer that, in the left case, the red pair are more similar than the rest, while when presenting the right distribution, most of them assert that the white apples are the most similar. In both cases, the most similar apples are identical. The key point is that the less likely the objects are (or the more they are specific), the more they are similar to themselves. This matches with Tversky's observation that specific features have more effect in similarity than generic features. The specificity of objects can be measured in terms of Information Quantity ($\mathcal{I}(\mathcal{X}) = -log(P(\mathcal{X}))$)

The second intuition is related with the fact that we can not dissect events into intersection or difference components. This assumption causes inconsistencies in Tversky's axioms with respect to observations when there are dependencies between the intersection and difference sets between objects. The **SIM axiom is grounded on the information quantities of single objects and the information quantity of their union**, that is, their joint probability. Notice that both the Pointwise Mutual Information $\left(\frac{P(x,y)}{P(x)\cdot P(y)}\right)$ and the conditional probability $\left(P(x|y) = \frac{P(x,y)}{P(y)}\right)$ are expressed in terms of joint and single probabilities. The SIM axiom states that:

Similarity Information Monotonicity Axiom: *If the Pointwise Information Quantity and the conditional probabilities between two objects grows, then their similarity grows. Formally if:*

$$\triangle PMI(\mathcal{X},\mathcal{Y}) \geq 0 \ \wedge \ \triangle P(\mathcal{X}|\mathcal{Y}) \geq 0 \ \wedge \ \triangle P(\mathcal{X}|\mathcal{Y}) \geq 0$$

Then $\triangle Sim(\mathcal{X},\mathcal{Y}) \geq 0$. *In addition, if at least one increase is strict then the similarity increase is strict.*

In other words, SIM basically states that Pointwise Mutual Information and conditional probabilities are the two basic dimensions of similarity, and similarity is monotonic with respect to both of them. If both grow, then the similarity grows. In the case of a trade-off between both, the similarity growth depends of the particular similarity metric.

SIM can be expressed in terms of increase of the joint and single information quantities (see Proof 2):

THEOREM 4.1. *SIM is equivalent to saying that there exists a positive similarity increase when both the single information quantity increase and their sum are higher than the joint information quantity increase:*

$$\triangle \mathcal{I}(\mathcal{X}) + \triangle \mathcal{I}(\mathcal{Y}) \geq \triangle \mathcal{I}(\mathcal{X} \cup \mathcal{Y}) \equiv \triangle PMI(\mathcal{X},\mathcal{Y}) \geq 0$$

$$\textit{and} \ \triangle \mathcal{I}(\mathcal{X}) \geq \triangle \mathcal{I}(\mathcal{X} \cup \mathcal{Y}) \equiv \triangle P(\mathcal{X}|\mathcal{Y}) \geq 0$$

$$\textit{and} \ \triangle \mathcal{I}(\mathcal{Y}) \geq \triangle \mathcal{I}(\mathcal{X} \cup \mathcal{Y}) \equiv \triangle P(\mathcal{X}|\mathcal{Y}) \geq 0$$

4.2 Formal Properties

The most important aspect of SIM is that it synthesizes all the proposed basic axioms. Proofs 3, 4, 5 and 6 in the additional material of this article prove that:

THEOREM 4.2. *Satisfying SIM is a sufficient condition to satisfy the identity, identity specificity, unexpectedness and dependency axioms.*

Given that SIM is defined in a symmetric manner, it can not be a sufficient condition for the Asymmetricity axiom. In fact, SIM does not say anything about the situation considered by the Asymmetricity axiom, given that the Pointwise Mutual Information does not change:

$$PMI(\mathcal{X}Y,\mathcal{X}) = PMI(\mathcal{X},\mathcal{X}Y)$$

and the conditional probabilities grow in opposite directions:

$$p(\mathcal{X}Y|\mathcal{X}) - p(\mathcal{X}|\mathcal{X}Y) = -\Big(p(\mathcal{X}|\mathcal{X}Y) - p(\mathcal{X}Y|\mathcal{X})\Big)$$

Therefore, the SIM conditions never hold.

Although we have discarded Tversky's axioms due to the need for considering dependencies between features, SIM has a direct correspondence with Tversky's Monotonicity axiom if we assume independence between intersection and difference set components. The following theorem states this (see Proof 7 in the additional material)

THEOREM 4.3. *Assuming independence between intersection and difference set components:*

$$\mathcal{I}(\mathcal{X} \cup \mathcal{Y}) = \mathcal{I}(\mathcal{X} \cap \mathcal{Y}) + \mathcal{I}(\mathcal{X} \setminus \mathcal{Y}) + \mathcal{I}(\mathcal{Y} \setminus \mathcal{X})$$

then the SIM axioms are equivalent to the following statement:

$$\triangle \mathcal{I}(\mathcal{X} \cap \mathcal{Y}) \geq 0 \ \textit{and} \ \triangle \mathcal{I}(\mathcal{X} \setminus \mathcal{Y}) \leq 0 \ \textit{and} \ \triangle \mathcal{I}(\mathcal{Y} \setminus \mathcal{X}) \leq 0$$

Note that the monotonicity axiom states that similarity grows when the intersection set grows nor the differences decrease. Given that adding elements to a set necessarily increases its information quantity:

$$\mathcal{I}(\mathcal{X}Y) \geq \mathcal{I}(\mathcal{X})$$

we can say that:

THEOREM 4.4. *Assuming independence between intersection and difference set components, satisfying SIM is a sufficient condition to satify Tversky's Monotonicity axiom.*

Going further, if we assume independence and equiprobability between features, then the information quantity of a feature set corresponds with its size ($I(\mathcal{X}) \propto |\mathcal{X}|$) and therefore:

THEOREM 4.5. *Assuming independence and equiprobability of features, SIM conditions are equivalent to stating that the intersection set size grows and the difference set size decrease:*

$$\triangle|\mathcal{X} \cap \mathcal{Y}| \geq 0 \ \textit{and} \ \triangle|\mathcal{X} \setminus \mathcal{Y}| \leq 0 \ \textit{and} \ \triangle|\mathcal{Y} \setminus \mathcal{X}| \leq 0$$

4.3 SIM as the Basis of Similarity

There are a number of reasons to believe that SIM could be the basic axiom of similarity:

(1) It synthesizes the proposed basic axioms (identity, identity specificity, unexpectedness and dependence).
(2) It models the traditional pointwise mutual information and the conditional probabilities as complementary components of similarity; and
(3) it has a direct correspondence with Tversky's axioms when assuming independence between intersection and difference components.

5 STUDY OF CURRENT SIMILARITY MODELS

In this section, we will see that, in general, current models do not satisfy our axioms, but they include additional mechanisms to mitigate this limitation.

5.1 Objects as Points in a Metric Space

Many similarity measures are grounded on the idea that object features can be represented as points in a metric space, which satisfies maximality, triangular inequality and symmetricity axioms. Typically, documents are modeled as vectors of word frequencies, and similarity is computed with metric distances such as euclidean or cosine.

Our *identity* axiom is satisfied, given that it is a relaxed version of maximality. However, maximality is not compatible with *identity specificity*: in metric spaces, every document is maximally similar to itself regardless of its specificity. Our unexpectedness axiom is also violated, given that features contribute to the overall similarity in a mutually-independent way.

In practice, this drawback of metric space models is mitigated by giving more weight to features with high specificity (or low likelihood). For instance, the cosine distance, which outperforms other measures such as the euclidean distance does not reward features if they are salient in both information pieces. For instance:

$$Cosine((2,10),(1,12)) = Cosine((2000,10),(1000,12))$$

The popular tf.idf feature projection function reduces the weight of words that are frequent in the collection. Notice that the second component $log\left(\frac{1}{p(w)}\right)$ of the tf.idf actually matches information quantity, assuming the document collection as sample space. Stopword removal also discards frequent features. However, *identity specificity* is not strictly satisfied, given that these models assume independence across features.

On the other hand, metric space measures do not consider the probabilistic dependency relationships between features. Therefore, the *dependency* and *unexpectedness* axioms are not satisfied. However, current approaches mitigate this problem at the representation level. One alternative is enriching texts by adding information from ontologies such as WordNet, or from related documents (pseudo-relevance feedback). Another approach, used by distributional semantics techniques such as Latent Semantic Indexing [7] and compositional distributional semantics [25], consists of collecting distributional information from the corpus and encoding it in high-dimensional vectors, obtaining a new representation of texts. The effect is that two texts with mutually dependent words will be enriched with similar information, thus increasing the overlap between them when computing similarity.

In this line, in recent years, neural networks have been used to map sentences into fixed-length vectors and then perform comparisons on these representations [23, 27]. In most cases, instead of applying metric distances, word similarity is estimated as the product of their corresponding vectors. Previous research has shown that the most popular representation algorithms (word2vec) converges into Pointwise Mutual Information (PMI) when computing the product of word vectors [2, 17]. ($\langle \vec{v_w}, \vec{v_{w'}} \rangle \approx PMI(w, w')$) PMI is analyzed in further sections, having other limitations in terms of axiomatics.

The compositionality of neural network language models for longer text pieces is still an open issue. In order to solve this issue, some of the most effective approaches to compute sentence similarity consist of combining this model with alignment mechanisms. These approaches are described in the following section. However linguistic units longer than words have been also projected into low dimensions as well as other features such as PoS tags and topic identifiers [33].

5.2 Objects and Transformations: Editing Distances

Another perspective consists of considering objects as things that can be transformed into other objects. In psychology, this corresponds with the *transformational approach* proposed by Hahn and Charter [11]. In the context of text processing, an example of editing based measure is WER, [26] which have been used to evaluate the performance of Machine Translation and Speech Recognition systems. An important strength of transformational models is that they are able to capture and align structures.

Some word alignment based approaches have achieved competitive results [12, 16, 21, 22] in the context of Semantic Textual Similarity tasks. The key point is that the lack of unexpectedness and dependency is mitigated by considering semantic distances between words (instead of substitutions): at the representation level, words are replaced by vectors of values in a reduced dimensionality space, or vectors of statistically related words. A measure like this outperformed 89 systems in the 2013 Semantic Textual Similarity shared task [1]. Some approaches also consider word order [18] and phrases [23].

However, there are two aspects that need to be solved. First, the alignment processes does not capture the dependencies between components in each object. The second one is that the similarity between structures is again an open issue. In principle, when assuming independence between transformational steps and structures, unexpectedness and dependency axioms can not be satisfied.

5.3 Texts as Feature Sets

Another family of similarity models follows the assumption that objects to be compared can be represented as sets of features. They are based on Tversky's axioms (matching, monotonicity and independence). One of the key contributions of Tversky is the *representation theorem* which states that similarity can be modeled as a linear function of the intersection and differences of sets; this is the Tversky linear contrast model:

$$Sim(\mathcal{X}, \mathcal{Y}) = \alpha_1 f(\mathcal{X} \cap \mathcal{Y}) - \alpha_2 f(\mathcal{X} \setminus \mathcal{Y}) - \alpha_3 f(\mathcal{Y} \setminus \mathcal{X})$$

Where f is a certain function which increases across subsumed sets ($f(\mathcal{X}) < f(\mathcal{X} \cup \mathcal{Y})$). This model fails to satisfy *expectedness* and *dependency* given that the difference component is assumed to be independent from the intersection between objects.

The parameterization (α_2 and α_3), on the other hand, captures *asymmetricity*, and the linear contrast model captures *identity specificity*, given that:

$$Sim(\mathcal{X}, \mathcal{X}) = \alpha_1 f(\mathcal{X} \cap \mathcal{X}) - \alpha_2 f(\mathcal{X} \setminus \mathcal{X}) - \alpha_3 f(\mathcal{X} \setminus \mathcal{X}) =$$

$$= \alpha_1 f(\mathcal{X}) - \alpha_2 f(\emptyset) - \alpha_3 f(\emptyset) = \alpha_1 f(\mathcal{X})$$

Therefore, self-similarity is not the same for every object. Assuming that f is related with the information quantity, *identity specificity* is satisfied.

Tversky studies showed that the parameterization depends on each particular scenario, and estimating the parameters is not straightforward. As an alternative, Tversky proposed the Ratio Contrast Model:

$$Sim(\mathcal{X}, \mathcal{Y}) = \frac{\alpha_1 f(\mathcal{X} \cap \mathcal{Y})}{\alpha_2 f(\mathcal{X} \setminus \mathcal{Y}) + \alpha_3 f(\mathcal{Y} \setminus \mathcal{X}) + \alpha_4 f(\mathcal{X} \cap \mathcal{Y})}$$

An advantage of this model is that it is easier to parameterize. Actually, whenever $\alpha_1 = \alpha_4$ the relative ordering between similarity instances values is not affected by the α_1 value, (See Proof 8). Therefore, only the relative value of α_2 and α_3 must be estimated in order to keep a consistent ordering between similarity values. The drawback for the ratio formulation is that *identity specificity* is no longer satisfied (see Proof 9 in the additional material).

Most set-based similarity measures can be derived from the ratio contrast model, taking the set size as salience function f. [20] contains a comprehensive description of these measures. Fixing different values for $\langle \alpha_1, \alpha_2, \alpha_3, \alpha_4 \rangle$ we obtain measures such as Jaccard, ($\langle 1, 1, 1, 1 \rangle$), Precision ($\langle 1, 1, 0, 1 \rangle$), Recall ($\langle 1, 0, 1, 1 \rangle$), Dice coefficiente ($\langle 2, 1, 1, 2 \rangle$), Anderberg coefficient ($\langle 1, 2, 2, 1 \rangle$) or First Kulczynski coefficient ($\langle 1, 1, 1, 0 \rangle$).

5.4 Objects as Sets of Events: Measures based on Information Theory

Other similarity models consider that object features are independent events with a certain information quantity and probability. In this line, Lin proposed a theoretical framework for similarity based on information theory [19]. Given a set of formal assumption, he obtains the *Similarity Theorem*, which states that: *The similarity between $\mathcal{X}$ and $\mathcal{Y}$ is measured by the ratio between the amount of information needed to state the commonality of $\mathcal{X}$ and calY and the information needed to fully describe what $\mathcal{X}$ and $\mathcal{Y}$ are:*

$$Sim(\mathcal{X},\mathcal{Y}) = \frac{logp(common(\mathcal{X},\mathcal{Y}))}{logp(description(\mathcal{X},\mathcal{Y}))}$$

where $common(\mathcal{X},\mathcal{Y})$ is a proposition that states the commonalities between them, and $common(\mathcal{X},\mathcal{Y})$ is a proposition that describes what $\mathcal{X}$ and $\mathcal{Y}$ are.

For string similarity, Lin proposed the following similarity model, considering words as independent features:

$$Lin(\mathcal{X},\mathcal{Y}) = \frac{2 \times \sum_{x \in \mathcal{X} \cap \mathcal{Y}} I(x)}{\sum_{x \in \mathcal{X}} I(x) + \sum_{y \in \mathcal{Y}} I(y)}$$

This expression matches Tversky's Ratio Contrast Model, if we use information quantity $(-\log P(\mathcal{X}))$ as f function [5]. Consequently, it inherits its limitations: it does not satisfy *identity specificity*, *unexpectedness* and *dependency*. In addition, assumption 4 (maximality) in Lin's work intrinsically contradicts the identity specificity constraint.

Cazzanti and Gupta [5], tried to improve Lin's distance by applying the linear contrast model with fixed parameters instead of the ratio.

$$RES = f(\mathcal{X} \cap \mathcal{Y}) - 0.5f(\mathcal{X} \setminus \mathcal{Y}) - 0.5f(\mathcal{Y} \setminus \mathcal{X})$$

where f salience function is the conditional entropy of random objects R regarding the observed features. $(f(\mathcal{X}) = H(R|\mathcal{X} \subset R))$. Basically, this salience function ensure that unfrequent features have more weight than frequent features. The interesting aspect of RES is that it captures identity specificity. However, it has the same limitations than Lin's distance in terms of unexpectedness and dependency. More explicitly , RES satisfies Tversky Monotonicity axiom (Property 8 in [5]) which is not compatible with the dependency axiom. As well as in the case of Lin distance, the underlying drawback is that the dependence between features in the differences and intersection is never considered. Regarding the asymmetricity axiom, these measures state fixed parameters that make the measure symmetric. There exists the possibility of tuning them for satisfying asymmetricity.

In general, the most interesting aspect of these models is that they are able to manage the specificity of features in the own similarity measure, instead of applying a previous feature projection functions such as tf.idf or stop-word removal.

5.5 Objects as Probabilistic Density Functions

In [6], Cha et al. describe 65 different similarity measures based on comparing probabilistic density functions. This perspective has a remarkable generalization power and, in fact, measures based on metric spaces and feature sets can be interpreted as density function similarities [6].

Again, a common drawback of all these measures is that they do not satisfy *unexpectedness* and *dependency*. The reason is modeling objects as probability distributions does not allow to infer statistical dependencies across objects.

In addition, none of them comply with *identity specificity*, because a distribution is equally similar to itself regardless of how much information it contains. Even measures based on Shannon's entropy [6] assign a maximal similarity (or minimal distance) to identical distributions. Consider the most paradigmatic measure, Kullback-Leibler divergence. Being P_i and Q_i the probability of the feature i in the object P or Q, their divergence is

$$d_{kl} \equiv \sum_i P_i \ln \frac{P_i}{Q_i}.$$

Now, if $P_i = Q_i$ for all i, then : $d_{kl} = \sum_i P_i \ln 1 = 0$. The same happens with other distribution entropy based measures such as Jeffreys, K divergence or Jensen-Shanon.

In summary, existing distribution based measures are not able to capture the identity specificity and dependency based axioms. In addition, although they are able to generalize geometric and set based measures, modeling similarity in this way does not allow to apply feature projection functions to mitigate these lacks.

5.6 Objects Generated by Probabilistic Distributions: Language Models

Another perspective consists of considering objects as single events generated by probabilistic distributions. Then, the similarity of objects is the likelihood of objects to be produced by the same probabilistic distribution. In text objects, this is the case of language models. In the basic language model approach proposed by Ponte and Croft [28] in the context of information retrieval, the similarity between a query and a document d is estimated as the probability that the query is produced by a probabilistic distribution θ_d inferred from the document d. $(Sim(Q,D) = p(Q|\theta_D))$ Assuming that θ_D is a multiple Bernouilli distribution:

$$p(Q|\theta_D) = \prod_{w \in Q} p(w|D) \prod_{w \notin Q} (1 - p(w|D))$$

where $p(w|D)$ is estimated as $\frac{freq(w,D)}{|D|}$.In practice, this requires a smoothing process in which the probability of unseen query words is estimated from the whole collection. Many improvements have been proposed since then. For instance, Hiemstra and Kraaij [13] and Miller et al. [24], proposed a variation based on multinomial word distributions.

In general, language models can satisfy *identity* and *identity specificity*. For instance, the last component in the model proposed by Zhai and Lafferty [37] is the sum of probabilities of query terms in the collection $(..+\sum_{w \in Q} p(w|C))$. This component is not considered given that it does not affect to the document ranking in a document retrieval task; but it would increase the self similarity of big queries as our axiom requires.

Strictly speaking, *unexpectedness* can not be satisfied: It is not possible to estimate the dependency between unseen query words and the document, given that the probability distribution is inferred from the document. However, according to the analysis in [37] and [36], the smoothing techniques have a connection with the idf effect,

and therefore they mitigate non-compliance with our *unexpectedness* axiom.

Dependency cannot be satisfied as well, given that there is no statistical dependence estimated from the document that connects different features from different objects. In addition, for computational reasons, language models in practice use multidimensional distributions that assume independence between features; therefore, the dependence within intersection and difference component sets is not considered. The use of n-grams (instead of single words) mitigates this problem.

5.7 Objects as Single Events in a Whole Probabilistic Distribution

The last approach consists of considering objects as single events in a global probabilistic distribution. From this perspective, in psychology, Shepard proposed to model similarity as the probability of one stimulus given another stimulus [31]: $Sim(\mathcal{X},\mathcal{Y}) = P(\mathcal{X}|\mathcal{Y})$.

The strength of the conditional probability as similarity model is that it trivially satisfies *dependency* and *unexpectedness*, given that adding different features to the second object can increase the estimated similarity. For instance:

P("Computer"|"Apple Desktop") > P("Computer"|"Apple").

The main limitation of conditional probability as similarity measure is that it does not comply with *identity specificity*, given that the self similarity is maximal and constant for every object ($P(\mathcal{X}|\mathcal{X}) = 1$).

The other well known approach is Pointwise Mutual Information (PMI), which is based on the idea that the more two objects are statistically independent, the less they are similar. PMI is estimated as:

$$PMI = \log\left(\frac{P(\mathcal{X},\mathcal{Y})}{P(\mathcal{X})\cdot P(\mathcal{Y})}\right)$$

which is zero when both events are independent. PMI can be also expressed in terms of information quantity:

$$PMI = \mathcal{I}(\mathcal{X}) + \mathcal{I}(\mathcal{Y}) - \mathcal{I}(\mathcal{X}\cup\mathcal{Y})$$

PMI have been used in multiple approaches to estimate pairwise word similarity. It also has been employed to predict concept similarity, by estimating their information quantities according to their depth in the hierarchical ontology.

As well as conditional probabilities, PMI satisfies *dependency* (see Proof 10). But the main strength of PMI is that, unlike the previous models, it captures the *identity specificity* case. In particular, the self similarity for any object corresponds with its *Information Quantity*:

$$PMI(\mathcal{X},\mathcal{X}) = \log\left(\frac{p(\mathcal{X},\mathcal{X})}{p(\mathcal{X})*p(\mathcal{X})}\right) = -log(p(\mathcal{X}))$$

The main lack of PMI is that (as its name suggests) it focuses only on the common features. For this reason, it cannot satisfy *unexpectedness* (See Prove 11)

5.8 Summary of Theoretical Analysis of Measures

Let us summarize the conclusions about this axiomatic analysis. Identity is satisfied by every model, but only measures based on the linear contrast ratio, language models and mutual information satisfy *identity specificity*. Most similarity models ignore this case, given that self similarities are not usually compared to each other in real scenarios. Only the conditional probability itself is able to satisfy *unexpectedness* in a strict manner.

However, this lack have been mitigated by techniques at representation level, such as text enrichment, pseudo relevance feedback or distributional semantics. The conditional probability also satisfies *dependence* at the cost of *identity specificity*. On the other hand, PMI is able to satisfy at the same time *dependence* and *identity specificity*, but not *unexpectedness*.

Finally, the asymmetricity is not the focus of many of the models. The reason is that, in most evaluation scenarios, the similarity ground truth annotated by humans for evaluation purposes is symmetric. In other scenarios, such as IR, texts (documents in a collection) are compared with one reference text (e.g query). Therefore, the asymmetric nature of similarity does not play a crucial role.

In general, we can extract the same conclusion than from the similarity information monotonicity axiom. Interpreting objects as single events in a whole distribution leads to conditional probability and Pointwise Mutual Information, which in combination are able to satisfy every axiom.

6 THE INFORMATION CONTRAST MODEL (ICM)

Assuming the SIM axiom as the core of similarity, we now derive the *Information Contrast Model*[1]. The SIM axiom suggests that a desirable measure should consider the relative increase of single, sum and union information quantities. That is, $\mathcal{I}(\mathcal{X})$ and $\mathcal{I}(\mathcal{Y})$ have a positive effect on similarity while $\mathcal{I}(\mathcal{X}\cup\mathcal{Y})$ has a negative effect.

Definition 6.1. The Information Contrast Model is the linear combination of the information quantity of each object and their union:

$$ICM_{\alpha_1,\alpha_2,\beta}(\mathcal{X},\mathcal{Y}) = \alpha_1 I(\mathcal{X}) + \alpha_2 I(\mathcal{Y}) - \beta I(\mathcal{X}\cup\mathcal{Y})$$

This measure can be interpreted as a generalized parametric version of Pointwise Mutual Information, being equivalent to:

$$ICM_{\alpha_1,\alpha_2,\beta}(\mathcal{X},\mathcal{Y}) = log\left(\frac{p(X\cup Y)^{\beta}}{p(X)^{\alpha_1}\cdot p(Y)^{\alpha_2}}\right)$$

6.1 Formal Properties

The most important property of ICM is that, it satisfies SIM under certain parameter ranges (see Proof 12):

THEOREM 6.2. *The information contrast model satisfies the similarity information monotonicity axiom when* $\alpha_1 + \alpha_2 > \beta > \alpha_1 > \alpha_2$.

The inequality $\alpha_1 > \alpha_2$ ensures that the measure is asymmetric, rewarding the similarity from the most specific text to the most general text. In a symmetric scenario, being $\alpha_1 = \alpha_2 = 1$, β must satisfy $2 > \beta > 1$. Note that, this theorem implies that ICM is able to capture all the axioms defined in this article.

ICM has a direct relationship with Pointwise Mutual Information and conditional probabilities depending of the parameters (See Proof

[1]We have selected this name by analogy with the Linear and Ratio Contrast model proposed by Tversky.

13):

$$\alpha \cdot \log(P(X|Y) * P(Y|X)) > ICM_{\alpha,\alpha,\beta} > \alpha \cdot \log(PMI(X,Y))$$

Therefore, being $\alpha_1 = \alpha_2 = \alpha$, the conditional probabilities and the mutual information profile the limits of ICM depending on the β value. In other words, **the mutual information and conditional probabilities are actually extreme cases of a generic measure.**

ICM is also closely related with set and information theory based measures. It is a generalization of the linear contrast model. In fact, assuming independence between component sets and information quantity as salience function, both ICM and the linear contrast model are equivalent.

$$ICM_{\alpha_1,\alpha_2,\beta}(\mathcal{X},\mathcal{Y}) =$$
$$(\alpha_1 + \alpha_2 - \beta)\mathcal{I}(\mathcal{X} \cap \mathcal{Y}) - (\beta - \alpha_1)(\mathcal{I}(\mathcal{X} \setminus \mathcal{Y})) - (\beta - \alpha_2)(\mathcal{I}(\mathcal{Y} \setminus \mathcal{X}))$$

In addition, there is a strong connection between ICM and language models as they are applied in IR. Language models estimate the similarity between text as the probability of the first text under the distribution derived from the second text.

$$ICM_{\alpha_1,\alpha_2,\beta}(q,d) = log\frac{P(q,d)^\beta}{P(q)^{\alpha_1} \cdot P(d)^{\alpha_2}}$$

Assuming a fixed $P(q)$ for all the retrieved documents, and $\alpha_2 = \beta$:

$$ICM_{\alpha_1,\alpha_2,\beta}(q,d) \propto log\frac{P(q|d)P(q)^\beta}{\cdot P(d)^{\alpha_2}} = \beta \log(P(q|d)) \propto P(q|d)$$

Therefore, according to ICM, language model in IR could be improved by considering the information quantity of candidate documents ($P(d)$) and a certain parameter.

7 EXPERIMENTAL PROOF OF CONCEPT

In this section, we focus on the counter examples that we used in Section 2 to invalidate, in the context of textual similarity, some of the axioms proposed elsewhere.

To do this, we need to estimate the information quantity of phrases such as *"Mickey Mouse"* or *"Apple desktop"*. We have used statistics from the Flickr search facility, which gives exact numbers (Web search engine statistics are larger, but search engines only offer approximate statistics on the number of hits and the way of determining the figures depends on the length of the query). For instance, for the word set *"Mickey apple"*, Flickr finds 2,141 documents. Given that Flickr stores around 13,000 million photos, it represents a probability of $0.164 * 10^{-6}$. We have made this estimation for every text in the examples, and we have computed ICM for each pair of texts. We have set ICM parameters as $\alpha_1 = 1.2, \alpha_2 = 1, \beta = 1.5$, arbitrary values that fit into the ranges stated in our theoretical analysis.

Table 1 shows the results. The first column contains the similarity inequality that we expect intuitively. The second and third columns contain the ICM values of the leftmost and rightmost text pairs in the inequality, and the last column checks if ICM agrees with our intuitions. The first example, for instance, shows that ICM assigns a higher self-similarity to "apple computer" than to "apple", in agreement with our identity specificity axiom and in disagreement with the maximality axiom from the state of the art. Overall, ICM satisfies all examples for our axioms, and violates previous axioms in the cases where they predict counter-intuitive results. This is of course anecdotal evidence rather than a quantitative confirmation of the validity of ICM, but it serves as a proof-of-concept of how ICM works.

Table 1: Capturing Counter Samples with ICM

Similarity instance comparison	ICM_1	ICM_2	
Counterexample for maximality axiom. Example for identity specificity axiom			
Sim("Apple Computer", "Apple Computer") > Sim("Apple', "Apple')	1.32	0.88	✔
Counterexample for symmetricity axiom			
Sim("North Korea", "China") > Sim("China", "North Korea")	2.86	-0.79	✔
Counterexample for dominance axiom			
Sim("Disney Mouse", "Game Mickey") > Sim("Disney **Mouse**", "Game **Mouse**")	1.32	0.88	✔
Counterexample for consistency and independency axioms (I)			
Sim("Mouse **Disney**", "Mickey **Disney**") > Sim("Mouse **Disney**", "Hardware **Disney**")	2.86	-0.79	✔
Counterexample for consistency and independency axioms (II)			
Sim("Mouse **Wireless**", "Hardware **Wireless**") > Sim("Mouse **Wireless**", "Mickey **Wireless**")	2.6	2.47	✔
Counterexample for monotonicity Axiom			
Sim("Apple Desktop", "Mouse Computer") > Sim("Apple", "Mouse")	4.03	-2.86	✔
Example for identity axiom			
Sim("**Apple Mouse**", "**Apple Mouse**") > Sim("Apple **Mouse**", "**Mouse**")	4.06	2.29	✔
Example for unexpectedness			
Sim("**Mickey**", "**Mickey** Mouse") > Sim("**Mickey**", "**Mickey** Apple")	2.59	1.51	✔

8 CONCLUSIONS

In this paper we have shown how axiomatic explanations of similarity from other fields (Tversky's axioms from the point of view of cognitive sciences, and metric spaces from the point of view of algebra) do not fit the problem of computing textual similarity, and we propose a new axiomatics. We have then performed a formal analysis of existing approaches to compute similarity, and we have seen that no previous model satisfies all our axioms. In many cases this is, however, mitigated with an enrichment at the representation level (for instance, idf weights add a notion of specificity into metric space distances which is not in the similarity model).

Our formal study leads us to introduce a new similarity model, the *Information Contrast Model*, which generalizes both Tversky's linear contrast model and Pointwise Mutual Information, and, unlike most existing similarity models, satisfies our axiomatic framework for a certain range of parameter values. In short, ICM states that similarity grows with the information quantity of individual objects (intuitively, rare object features that tend to occur simultaneously are strongly connected), and decreases with the information quantity of the union (intuitively, object features that rarely occur simultaneously have little connection). We have presented a small proof of concept over Flickr statistics, where ICM satisfies the predictions of our axiom set, and aligns with our intuition in the cases where previous axioms prescribe counterintuitive behavior.

The main challenge derived from our study is how to estimate properly the information quantity (or probability of feature sets) without assuming feature independence.

REFERENCES

[1] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, USA, 32–43.

[2] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A Latent Variable Model Approach to PMI-based Word Embeddings. *TACL* 4 (2016), 385–399.

[3] E. G. Ashby and N. A. Perrin. 1988. Toward a unified theory of similarity and recognition. *Psychological review* 95, 1 (1988), 124–150.

[4] Fred Attneave. 1950. Dimensions of similarity. *American Journal of Psychology* 63, 4 (1950), 516–556.

[5] Luca Cazzanti and Maya Gupta. 2006. Information-theoretic and Set-theoretic Similarity. In *2006 IEEE International Symposium on Information Theory*. IEEE, 1836–1840.

[6] Sung-Hyuk Cha. 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. (2007).

[7] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41, 6 (1990), 391–407.

[8] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, New York, NY, USA, 49–56.

[9] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, New York, NY, USA, 480–487.

[10] R. L. Goldstone, D. L. Medin, and D. Gentner. 1991. Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology* 23 (1991), 222–264.

[11] Ulrike Hahn, Nick Chater, and Lucy B. Richardson. 2003. Similarity as transformation. *Cognition* 87, 1 (Feb. 2003), 1–32.

[12] Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics.

[13] D. Hiemstra and W. Kraaij. 1998. Twenty-One at TREC-7: ad-hoc and cross-language track. In *Seventh Text REtrieval Conference, TREC 1998 (NIST Special Publications)*, E.M. Voorhees and D.K. Harman (Eds.), Vol. 500-24. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 227–238.

[14] C. Krumhansl. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review* 85, 5 (1978), 445–463.

[15] J. Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1 (1964), 1–27. http://EconPapers.repec.org/RePEc:spr:psycho:v:29:y:1964:i:1:p:1-27

[16] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, David Blei and Francis Bach (Eds.). JMLR Workshop and Conference Proceedings, 957–966. http://jmlr.org/proceedings/papers/v37/kusnerb15.pdf

[17] Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding As Implicit Matrix Factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2177–2185. http://dl.acm.org/citation.cfm?id=2969033.2969070

[18] Lin Li, Xia Hu, Bi yun Hu, Jun Wang, and Yi ming Zhou. 2009. Measuring Sentence Similarity from Different Aspects. In *In Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*. 2244–2249.

[19] Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 296–304.

[20] Maurice HT Ling. 2010. Distance Coefficients between Two Lists or Sets. (2010).

[21] Rahul Malik, L. Venkata Subramaniam, and Saroj Kaushik. 2007. Automatically Selecting Answer Templates to Respond to Customer Emails.. In *IJCAI*, Manuela M. Veloso (Ed.). 1659–1664.

[22] R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*. Boston, Massachusetts.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546 (2013). http://arxiv.org/abs/1310.4546

[24] David R. H. Miller, Tim Leek, and Richard M. Schwartz. 1999. A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 214–221.

[25] J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. *proceedings of ACL-08: HLT* (2008), 236–244.

[26] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*.

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543.

[28] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM, New York, NY, USA, 275–281.

[29] E. Rosch. 1975. Cognitive reference points. *Cognitive Psychology* 7 (1975), 532–547.

[30] E. Z. Rothkopf. 1957. A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology* 53 (1957), 94–101.

[31] R.N. Shepard. 1987. Toward a universal law of generalization for psychological science. *Science* 237 (1987), 1317–1323.

[32] Roger N Shepard. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* 27, 2 (1962), 125–140.

[33] Yangyang Shi, Pascal Wiggers, and Catholijn M. Jonker. 2012. Towards Recurrent Neural Networks Language Models with Linguistic and Contextual Features. ISCA, 1664–1667.

[34] A. Tversky. 1977. Features of similarity. *Psychological Review* 84 (1977), 327–352.

[35] Amos Tversky and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological review* 89, 2 (1982), 123.

[36] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (March 2008), 137–213.

[37] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214.