A New Measurement of Systematic Similarity

Yi Guan, *Member, IEEE*, Xiaolong Wang, and Qiang Wang

***Abstract*—The relationship of similarity may be the most universal relationship that exists between every two objects in either the material world or the mental world. Although similarity modeling has been the focus of cognitive science for decades, many theoretical and realistic issues are still under controversy. In this paper, a new theoretical framework that conforms to the nature of similarity and incorporates the current similarity models into a universal model is presented. The new model, i.e., the systematic similarity model, which is inspired by the contrast model of similarity and structure mapping theory in cognitive psychology, is the universal similarity measurement that has many potential applications in text, image, or video retrieval. The text relevance ranking experiments undertaken in this research tentatively show the validity of the new model.**

***Index Terms*—Matching objects pair, structure mapping theory, systematic similarity, systematic similarity measurement criterion, systematic similarity model (SSM).**

I. Introduction

SIMILARITY is a binary relationship that exists between conceptual and perceptual objects. G. W. Leibniz once argued that no two leaves could ever be exactly the same; but in the sense of ontology, any two objects in the world show some degree of similarity. It can readily be argued that the relationship of similarity may be the most universal relationship that exists between any two objects in either the material world or the mental world.

Similarity plays a foundational role in cognition theories. Many cognitive processes, such as problem solving, categorization, memory retrieval, inductive reasoning, etc., require a good understanding of the methodology that involves similarity assessment. Therefore, the measurement of similarity has been the focus of cognitive science for decades. Meanwhile, the needs for the retrieval of video, image, audio, and text data from mass data revive the people's interest in this subject with the rapidly growing amount of information available on the Internet. The research on similarity measurement for content-based text/image/video retrieval has become one of the active research issues over the last decade [1]–[4].

Similarity can be classified into different categories from different perspectives. Gentner has distinguished six kinds of similarity (i.e., literal similarity, analogy, abstraction, metaphor, anomaly, and mere appearance) based on the kinds of predicates shared [5]. According to the source of compared objects, similarity can be categorized into conceptual similarity and perceptual similarity. Similarity can also be divided into dimensional similarity and global similarity [6]. This distinction essentially focuses on the fact that one can perceive two objects as of similarity in terms of some holistic perception of them or by only along certain discriminable dimensions. Global similarity, which is an overall literal similarity in Gentner's taxonomy, will be the focus of this paper. However, the term "systematic similarity" will be the substitute of "global similarity" to emphasize that the objects compared are richly structured, or, in other words, each object is hierarchically constituted of subobjects linked by heterogeneous relations. This paper deals with the measurement of systematic similarity between two different objects.

Despite its importance and universality, similarity measurement theory is not well established in that many theoretical and realistic issues are still under controversy. Such a situation results from the fact that similarity is a hybrid concept across many branches of cognitive science, and that similarity degree is considerably sensible and variable to be tested or validated as a psychological measurement. However, it is of great value to propose a new theoretical framework of similarity measurement that not only conforms to the nature of similarity but also enjoys the following advantages.

1) *Consistency*. A rational framework should be coincident with the existing experimental facts, common senses, and practical experiences. Years of studies of the similarity models and features have witnessed many psychological experiments through which some features of similarity have already been revealed and widely accepted [7]–[10]. The new model presented in this paper has to follow the valuable conclusions of that research. On the other hand, some models of similarity are criticized for their violation of the human's common sense [11], and such drawbacks should be revised by the new model.
2) *Conciseness.* A rational framework should have few (even no) prior conditions or hypothesis. A. Einstein once said, "The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms." Hence, unnecessary hypotheses or prior conditions will lead to an unstable theoretical foundation of the whole theory.
3) *Compatibility*. A rational framework should stem from current theories, discard the dross, and select the essential (hence, a new integrated model). The current models of similarity have already been successfully applied in many fields, but the pursuits for a more precise and efficient measurement have never ceased. The new model must

Manuscript received January 18, 2006; revised November 10, 2006. This work was supported by the National Natural Science Foundation of China under Grant 60504021 and Grant 60736044. This paper was recommended by Associate Editor J. Miller.

Y. Guan and X. Wang are with the Harbin Institute of Technology, Harbin 150001, China (e-mail: guanyi@hit.edu.cn; wangxl@insun.hit.edu.cn).

Q. Wang is with the New Search Research and Development Department, Baidu Company, Beijing 100080, China (email: wangqiang@baidu.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2008.918611

be compatible with all the reasonable ingredients of the existing similarity models.

4) *Simplicity.* A rational framework should be as simple as possible. L. Tolstoy once observed that "Every important idea is simple," and the purpose of modeling a concept as important as similarity leads to a presumption of the framework to be of simplicity.
5) *Universality*. Although it is controversial that there exists a universal measurement of similarity, a similarity model could achieve sufficient universality if it is based on modeling human similarity assessment, and the objects are represented by a set of subobjects linked by different interactions and relationships.

The remainder of this paper is organized as follows. Section II surveys the current psychological models of similarity. Section III presents theoretical foundations and essential preconditions of systematic similarity measurement. The measurement theory of systematic similarity is established in Section IV, whereas a new model of information retrieval is discussed in Section V as one application of the theory. In Section VI, these ideas are tested by text relevance ranking experiments. Related works are discussed in Section VII, and conclusive affirmations are presented in Section VIII.

II. Current Models of Similarity

Similarity can be measured by a wide variety of methods. Generally speaking, there are four major psychological models of similarity, i.e., geometric, featural, alignment based, and transformational models.

A. Geometric Model

One of the most influential similarity models is the geometric model [8], which is exemplified by multidimensional scaling (MDS) models [12]. In MDS models, each object is represented by a point in a multidimensional feature space, and similarity is then inversely related to the distance between points in the space. A Euclidean metric or a city block metric is often used for distance measurement, which is regarded as a reasonable conformation to human similarity judgments.

The geometric models defaultly presume minimality $(D(A,B) \geq D(A,A) = 0)$, symmetry $(D(A,B) = D(B,A))$, and triangle inequality $(D(A,B) + D(B,C) \geq D(A,C))$. Unfortunately, violations of all three assumptions are empirically observed [7]. Although the geometric models can be modified to correct these assumptions [13], the dimensions in MDS are assumed to be independent, which also seems inappropriate with the experimental data [14].

B. Feature Contrast Model

Tversky and Gati proposed the feature contrast model, wherein similarity is determined by common and distinctive features of the objects compared [15]. The similarity degree of A and B, i.e., $S(A,B)$, is expressed as a linear combination of the measure of common and distinctive features: $S(A,B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$. The term $A \cap B$ represents the features that items A and B have in common. $A - B$ represents the features that A enjoys but B does not. $B - A$ represents the features that B possesses but A does not. The terms θ, α, and β reflect the weights given to the common and distinctive components, and the function f is often assumed to be additive.

The contrast model can account for the asymmetries and the invalidations of triangle inequality that occur in similarity judgments [15]. However, in this model, the objects are represented by a set of features, and the common features may not increase the similarity degree unless they belong to the aligned structures of the entities compared.

C. Alignment-Based Model

Both geometric and featural models might fail to compare things that are richly structured. However, a hierarchical representation of things is the most efficient. In such cases, a comparison of the entities involves not only simply matching features but also determining which components correspond to, or align with, one another. In alignment-based models, the matching features influence the similarity more if they belong to the parts that are placed in correspondence, and parts tend to be placed in correspondence if they have many features in common and if they are consistent with other emerging correspondences [16], [17].

There are many evidences that show the rationality of the alignment-based model [10]. Nowadays, the growing attention of research communities has been paid to similarity measurements on tree-structured data [18], [19]; however, much of this research is devoted to domain-specific measures. Until now, no analytic expression of a universal structural alignment-based similarity measurement has been proposed, which will be the focus of this paper.

D. Transformational Model

Another typical similarity model is based on transformational distance. The similarity of two objects is assumed to be inversely proportional to the number of operations required to transform one entity so as to be identical to the other [20], [21].

Although structured representations of entities are within the reach of the transformational model, the entities are represented by Kolmogorov complexity [21], which restricts its usability.

All the four approaches have enjoyed some success in quantitatively predicting the people's similarity assessments. Testing and analyzing these models' approaches have been a major ongoing subject of research. In addition to these investigations, many domain-specific similarity models are also presented [2], [3], [22]. Miscellaneous definitions, notions, hypotheses, or approaches confuse our understanding on the nature of similarity. A unified model of similarity is needed to end the endless controversies and proposals.

In this paper, we present a unified theoretical model, namely, the systematic similarity model (SSM). The new model not only extracts valid ingredients from current similarity models and experimental evidences but also manifests itself in a simple mathematical form. Since it is based on the simulation of human similarity assessment and independent of cognitive objectives, the model has potential for becoming a general-purpose model of similarity.

III. Foundations of Systematic Similarity Measurement

A. Structure Mapping Theory

Structure mapping theory for analogy, which describes the mechanism of human analogical processing [23], has achieved considerable success in psychology and computer science, and, thus, has been considered as "unquestionably the most influential work to date of the modeling of analogy-making" [24]. Although analogy is only a kind of similarity in Gentner's taxonomy, since "Similarity comparisons (like analogical comparisons) require structural alignment" [25], we choose the structure mapping theory as the theoretical foundation for systematic similarity measurement.

The basic idea of structure mapping theory is that an analogy (and "systematic similarity" hereafter) comparison contains a process of mapping knowledge from the base object into the target object that conveys that a system of relations that holds among the base objects also holds among the target objects. According to Gentner [5], [23], during similarity evaluation, the match between two structured representations must be structurally consistent, i.e., it must conform to the one-to-one mapping and connectivity constraints (or parallel connectivity). One-to-one mapping means that for any given match between representations, each object in one representation will map to at most one object in the other representation. Connectivity constraints mandate that if a match is made between representations, the subobjects of those representations must match as well [10]. These cognitive processes have been proved by many experimental evidences [10], [16], [17].

Central to the structure mapping process is the principle of systematicity: people prefer to map systems of objects that contain higher-order relations rather than to map-isolated objects [23]. Other works have demonstrated that subjects often find structural matches more compelling than feature matches [26], [27]. These discoveries imply that the human's tacit preference in similarity match is from higher-order relation down to lower-order relation, then to entity match, and finally to attribute match. The SSM has to follow such features of human cognition.

Structure mapping theory deeply impacts on the investigation of the human cognitive model and arouses much attention in psychology, computer science, and artificial intelligence. A well-known implementation of the psychological theories is the Structure Mapping Engine (SME) [28], [29].

B. SME

The SME is a program built to explore the computational aspects of Gentner's structure mapping theory of analogical processing. Given the descriptions of a base and a target, SME constructs all the structurally consistent mappings between them. The mappings consist of pairwise matches between statements and entities in the base and target, plus the set of analogical inferences. SME also provides a global structural evaluation score based on local structural matches.

SME takes two arguments, which are called the source and the target, respectively. The arguments are represented by description groups that contain a list of entities and predicates. Entities represent objects or concepts, whereas predicates are one of relation, attribute, and function. Predicates are a general way to express knowledge for SME. Relation predicates contain multiple arguments that can be other predicates or entities. For example, a relation "GO(agent, from, to)" is named "GO" and takes three arguments, i.e., "agent", "from", and "to". The attribute predicates are the properties of an entity. An example of an attribute is (RED apple), which means that apple has the attribute of red color. Finally, function predicates map an entity into another entity or constant.

Given the descriptions of a source and a target, the SME algorithm returns a structural match evaluation score (which has identical function as similarity degree) following the four main steps:

1) local match construction: finds all pairs of entities or predicates that can potentially match;
2) global mapping construction: combines the local matches into maximal matching predicates;
3) candidate inference construction: derives the inferences suggested by each gmap;
4) match evaluation: attaches evidence to each local match and uses this evidence to compute structural evaluation scores for each gmap until the final structural match evaluation score is obtained.

Three salient characteristics of the algorithm are worthy of being emphasized. 1) The algorithm formulates and implements a set of predicate calculus to represent objects in order to ensure domain independence of the engine. 2) The local-to-global alignment process demonstrates the bottom-up manner of the structural evaluation score computing process. 3) The algorithm uses "evidential weight" to quantify matching items and derives the structural evaluation score from local match scores. As will be illustrated later, all these characteristics are inherited by a systematic similarity measurement algorithm.

C. Outline of Systematic Similarity Measurement

In [28], the analogical processing is decomposed into three stages.

1) Access: The source situation is retrieved with regard to the given target situation in this stage.
2) Mapping and inference: That is, structure mappings are carried out between the source and the target.
3) Evaluation and use: In this stage, the quality of the match is estimated.

Likewise, the systematic similarity measurement process of the human being can be divided into three stages, i.e., assessment, mapping, and evaluation. People determine the target object and the source object, which will be under similarity evaluation in the assessment stage. In the mapping stage, people try to map the higher-order structure between two objects. Because the higher-order mapping can only be implemented through lower-order structure mappings, the mapping process

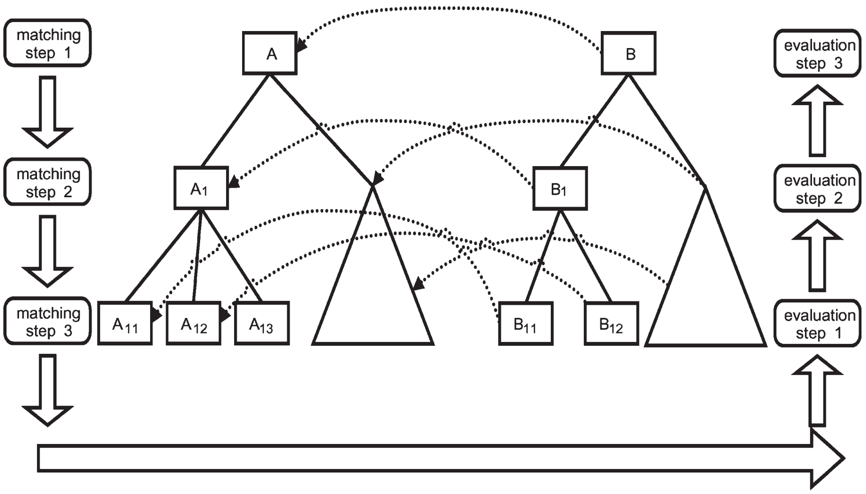


Fig. 1. Process of systematic similarity evaluation.

will proceed in a top-down manner through the route of higher-order relation → lower-order relation → entity match → attribute match, following the principle of systematicity. After that, the evaluation stage will start like the local-to-global alignment process of SME. This process will be along the reverse route to hierarchically achieve matching scores between attributes, entities, lower-order relations, and higher-order relations, until the overall similarity score between the two objects is finally gained. An example in Fig. 1 illustrates the mapping and evaluation process.

D. Prerequisites of SSM

The essence of SSM is inspired from a revision of Tversky's contrast model in which the similarity of compared objects is assumed to increase with its common features and decrease with its different features. Instead, in SSM, the similarity degree of the compared objects increases with their similar (matched) subobjects and decreases with their different (unmatched) subobjects $(^*)$, which is named as "systematic similarity measurement criterion" in this paper. The former part of the criterion is motivated by the structure mapping theory [23]. On the other hand, the latter part of the criterion is inspired by the evidences that "structural differences in the form of distinctive global properties decrease the similarity of a pair of items" [25]. To estimate the systematic similarity degree of two objects, SSM will always distinguish one-to-one aligned subobjects from unmatched subobjects, and then the similarity degree will be computed under the criterion $(^*)$. As will be illustrated later, this process can be performed in a simple and recursive way if the range of similarity degree is constrained in [0, 1]. A comparison study of Tversky's hypothesis and systematic similarity measurement criterion inspires us that the criterion can be seen as the discrete version of Tversky's hypothesis. Or put another way, Tversky's hypothesis can be seen as the continuum version of the criterion. The two hypotheses could be exactly the same if the attributes are regarded as objects of continuum structure, whereas the entities and higher-order structures are regarded as objects of discrete structure. This is why we treat attribute as a special kind of object in this paper.

A number of proposed similarity measures explain similarity as a distance in some feature space that is assumed to be a metric space [8], [12]. Such measures redundantly attach properties such as symmetry and triangle inequality to the similarity measure [15]. Indeed, many counterexamples of these assumptions have already been pointed out so that it is reasonable to discard these unnecessary preconditions. In SSM, besides the systematic similarity measurement criterion, only one prerequisite survives, which is called "range prerequisite" of similarity: "for objects A and B, if and only if A and B are the same, in other words, A is copy of B or vice versa, then their similarity is 1, otherwise, their similarity is a scalar value in range [0,1)."

The three intuitions about similarity measurement are proposed in [36]: "Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are. Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are. Intuition 3: The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share." The prerequisites of SSM are obviously in conformity with the three intuitions.

Some researchers argue that complex objects that are identical seem more similar to each other than simpler objects that are identical [15]; in the author's opinion, however, it may due to an illusion of sense.

IV. SSM

Everything in the world can be expressed as a system, i.e., a set of objects linked by different interactions and relationships. Systematic similarity is a global one caused by the similarity of all the respective correlative objects or features of compared items, not merely some specific features or relations. For example, an analogy of atom and the solar system is not systematic similarity, because only similarity in orbit exists: electrons orbit the nucleus, whereas planets orbit the sun. From the viewpoint of systematic similarity, they are completely different. An SSM is defined as a binary tuple $[F, S]$, where

F is a framework for modeling the systematic structures, and $S(A, B)$ is a binary function that computes the systematic similarity degree of object A and object B, which have been modeled by F.

A. Representation Conventions for Systematic Structures

On the basis of a high-order predicate argument structure, we propose the object representation conventions in SSM for the sake of maximizing the generality of the model. These conventions are mainly taken from [5], which Gentner uses to describe her idea.

An object is the core concept of representation framework in SSM. It simply means the thing under similarity evaluation, which could be either conceptual or perceptual. The two arguments of SSM are called source object and target object, respectively.

An object is composed of three types, i.e., attribute, entity, and relation.

1) Attribute. An attribute describes some property of an entity. In this paper, the attributes are represented by words in lowercase letters, for example, "red," "fast," etc.
2) Entity. An entity is the most basic object or concept of a discrete structure. In this paper, the entities are declared as words in uppercase letters with a list of attributes as follows:

$$\text{ENTITY\_NAME}(\text{attribute1}, \text{attribute2}, \ldots, \text{attribute } n).$$

   For example, APPLE(red, big, sweet) describes a red, big, and sweet apple.
3) Relation. A relation links a set of relations or entities to form a higher-order structure. As the most complicated object, a relation stipulates not only the number and order of its arguments but also the semantic role of each argument. For example, the relation "GIVE(I, APPLE, YOU)" describes an ordered triple relation in which all the parameters are agent (with the instance "I"), object(with the instance "APPLE"), and objective(with the instance "YOU"), respectively. However, they are not marked by the representation conventions of this paper for conciseness of writing.

Relations are declared as words in uppercase letters with a list of relations or entities as RELATION_NAME $(\text{RELATION } 1|\text{ENTITY } 1, \text{RELATION } 2|\text{ENTITY } 2, \ldots, \text{RELATION } n|\text{ENTITY } n)$.

Domains and situations are viewed as the systems of object attributes, objects, and relations between objects in [5]. These items correspond with the attribute, entity, and relation of this paper, respectively. The only difference between them is that the attributes are considered as a unary relation in [5], but in this paper, the attributes are treated as objects of the continuum system.

Given the set of object convention framework, the object in Fig. 2 can be described as SMILEY(FACE(circular, orange), EYES(LEFT_EYE(elliptic, brown), RIGHT_EYE(elliptic, brown)), MOUTH(LIP(curved, brown), TEETH(TOOTH1 (white), . . ., TOOTH12(white)))).

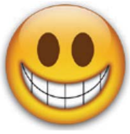


Fig. 2. Smiley.

B. Computation of Systematic Similarity Degree

Although the computation of systematic similarity degree requires all the respective correlative objects or features of the compared items, it does not mean that all containing objects or features have to be enumerated before measuring takes place, and in most cases, it is even impossible to do so. Instead, a main factor analysis step, as many similarity measurement researches contain [3], [30], [31], is firstly needed.

The following discussion will center on an example that is named as the *sentence match example*. The two sentences are as follows.

1) Target sentence: "And though I bestow all my goods to feed the poor, and though I give my body to be burned, and have not charity, it profiteth me nothing."
2) Source sentence: "If I give everything I own to the poor and even go to the stake to be burned as a martyr, but I do not love, I've gotten nowhere"[1] [32].

A complete computation process of the systematic similarity degree will be illustrated between the two sentences. It should be noted that SSM is a domain-independent model that does not constrain itself to computational linguistics. We use this example simply to clarify our discussions. The issue will be addressed by the end of the fourth part.

1) Main Factors:

Entity Similarity Degree: Just as an entity is the most basic constituent of a discrete system that cannot be divided further, the entity similarity degree is the most basic factor for systematic similarity measurement.

According to the range prerequisite of similarity, the entity similarity degree is a value in the range [0, 1], which is denoted by μ. As the simplest and initial systematic similarity, μ has to be acquired from common or distinctive features of compared entities since no structural alignment process can be performed. For this purpose, featural models of similarity, including contrast models or geometric models, are still applicable.

Words are entities in the sentence match example; therefore, the entity similarity degrees are the semantic similarity degrees between words. Many methods have been proposed for the computation of word similarities, for example, information-content-based methods [56], distance (minimum distance, Euclid distance, Manhattan distance)-based methods [33], statistical-distribution-based methods [30], and methods based on the integration of multiple information sources [34]. Lin summarized the current similarity measuring methods and presented an information-theoretic definition of similarity [36]. In

[1] The target sentence is quoted from "The Message," and the source sentence is quoted from the King James Version of the Bible. This example is quoted from [32].

TABLE I
ENTITY SIMILARITY DEGREE OF THE SENTENCE MATCH EXAMPLE

Target Source	And	though	I	bestow	all	…
If	-	0.8	-	-	-	…
I	-	-	1.0	-	-	…
give	-	-	-	0.9	-	…
everything	-	-	-	-	0.5	…
I	-	-	1.0	-	-	…
…	…	…	…	…	…	…

In this example, entity similarity degrees are arbitrarily set to make sense

the sentence match example, entity similar pairs and correspondent entity similarity degrees may look like that of Table I.

MOPs and associated similarity degree: According to Gentner's theory [5], [23], during similarity degree evaluation, a complete match between two structured representations is fulfilled, and such a match must conform to the "one-to-one mapping" and connectivity constraints.

The one-to-one mapping process can be modeled by "Matching objects pair." When object A and object B are under similarity evaluation, for a subobject A_i of object A, a subobject B_j of object B must satisfy two conditions to become the one-to-one mapping object of A_i. 1) It must satisfy some semantic role constrains or order constrains required by the relations. For example, the word "I" of the target sentence in Table I has two matching candidates in the source sentence with maximum similarity degree. To decide the right "I", the semantic role and the order of the relation will help to choose the first one because both of them are agents of respective relations. 2) The similarity degree between A_i and B_j must be as great as possible and at least greater than some predefined threshold μ_0. If B_j satisfies both condition 1 and condition 2, then $\langle A_i, B_j \rangle$ forms a *matching objects pair* with similarity degree μ. Under this paradigm, the *structural mapping process* during similarity estimation becomes that of finding all the matching object pairs (MOPs) between the given objects.

Condition 2 shows that the similarity degree of objects is derived from the MOPs between their subobjects and associated similarity degrees; therefore, the connectivity constraints are also satisfied under the definition of MOP.

In the sentence match example, some higher-order MOPs may be listed as follows (Table II).

Weight of objects: Different objects have different effects on the systematic similarity degree, as is quantified by weight, which represents the importance of an object with regard to its contribution to the systematic similarity. For example, in an exploration of the resemblance of human faces, the eyes may be assumed to be of higher importance than other facial organs.

The weight of an entity can be considered as a function of its attributes. The weight of a relation may be considered as a linear combination of the weights of its subobjects. For example, the weight of a relation is set to be the maximum weight of its subobjects in an implemented version of SSM. One may think that weight is constant for a given entity; however, in many scientific fields, weight is vulnerable to change with context, perspective, and preference, which causes the mutability of a similarity degree judgment. For example, in natural language, one word may have multiple meanings, corresponding to multiple weights in which only one meaning is active in some specific context, which is known as word sense disambiguation.

TABLE II
MOPS IN THE SENTENCE MATCH EXAMPLE

Sentence Order No.	Target	Source
4	And though I bestow all my goods to feed the poor, and though I give my body to be burned and have not charity	If I give everything I own to the poor and even go to the stake to be burned as a martyr but I don't love
	it profiteth me nothing	I've gotten nowhere
3	And though I bestow all my goods to feed the poor	If I give everything I own to the poor
	and though I give my body to be burned	and even go to the stake to be burned as a martyr
	and have not charity	but I don't love
…	…	…

In practice, weight is usually quantified by some combination of statistical measure. For example, a variety of term weighting schemes have been explored for information retrieval. The most successful and widely used scheme is the "term frequency * inverse document frequency" weighting scheme, which is commonly abbreviated as "tf*idf." The measure has many variations; some of them are summarized in [35].

In SSM, weight is a relative concept that links the subjective world and the objective world. The mutability of the weight may be one of the reasons why misunderstanding takes place, i.e., that different processes for assessing similarity are probably used for different tasks, domains, and stimuli [11]. (The other factors include the choice of features, structural descriptions, etc.). In addition, it also brings about more subjectivity and difficulty for measuring the systematic similarity degree.

Systematic structure: Systematic structure is the name of the overall attributes, entities, and relations, and their organizational forms for a given object. Attributes are objects of continuum structure, entities are the most basic objects of discrete structure, and relations are compound objects of discrete structure.

The influences of systematic structure on systematic similarity measurement are of two aspects, i.e., vertical and horizontal influence (Fig. 3). Vertical influence is determined by the principle of systematicity, which prescribes that the order of structure mapping is along the route from higher-order structure down to lower-order structure until attribute match. On the other hand, horizontal influence is determined by the systematic similarity measurement criterion with matched and unmatched objects during structural alignment, which is an inference of structure mapping theory.

2) Systematic Similarity Computing Function: According to the above analysis, the magnitude of the systematic similarity between two objects is related to the following main factors:

1) systematic structure of the compared objects;

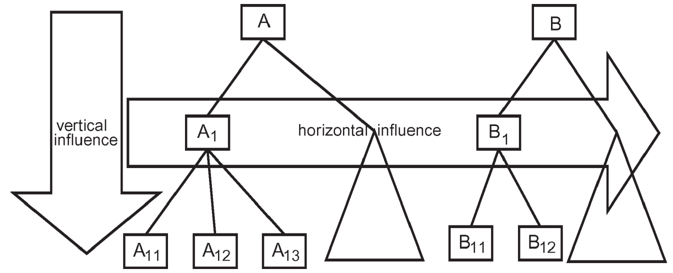


Fig. 3. Influences of systematic structure on systematic similarity measurement.

2) weight of the objects;
3) matching object pairs and associated similarity degrees (including the MOPs between entities and entity similarity degrees).

Other factors fall into these main factors, e.g., the categories of subobjects lie in systematic structures and weights.

Given an object $O_A(A_1, A_2, \ldots, A_m)$, let $A = \{A_1, A_2, \ldots, A_m\}$, $m = |A|$; and given an object $O_B(B_1, B_2, \ldots, B_n)$, let $B = \{B_1, B_2, \ldots, B_n\}$, $n = |B|$; here, $A_i (1 \leq i \leq m)$ and $B_j (1 \leq j \leq n)$ are subobjects. Letting $\mu = similarity(A_i, B_j)$ represent the similarity degree of subobjects A_i and B_j, and stipulating that $0 \leq \mu \leq 1$, $\mu = 1$ iff $A_i \equiv B_j$. A MOP is denoted by s_i with similarity degree μ_i. All the MOPs are constructed by the structural mapping process. Provided that the number of MOPs is p, and letting $s_1, s_2, \ldots, s_p \in A \times B$ denote the MOPs of objects A and B, their similarity degrees are $\mu_1, \mu_2, \ldots, \mu_p$, respectively, $\mu_i \geq \mu_0 (i = 1, \ldots, p)$. The weight of subobject A_i is denoted by $d(A_i)$. If the systematic similarity degree is $SS(O_A, O_B)$, then

$$SS(O_A, O_B) = f\left(m, n, p, \mu_1, \mu_2, \ldots, \mu_p, d(A_1), d(A_2), \ldots, d(A_m), d(B_1), d(B_2), \ldots, d(B_n)\right). \quad (1)$$

$SS(O_A, O_B)$ is a multivariate function that meets the systematic similarity measurement criterion.

Because each subobject is represented by its weight during systematic similarity estimation, the systematic similarity measurement criterion is decomposed into the following five monotonicity conditions. Hence, the systematic similarity between two objects is as follows:

1) an increasing function of weights of the subobjects in MOPs;
2) a decreasing function of weights of the subobjects that are not in MOPs;
3) an increasing function of the similarity degrees of MOPs;
4) an increasing function of the number of MOPs;
5) a decreasing function of the number of subobjects that are not in MOPs.

C. SSM

Given an object $O_A(A_1, A_2, \ldots, A_m)$, $A = \{A_1, A_2, \ldots, A_m\}$, $m = |A|$, and an object $O_B(B_1, B_2, \ldots, B_n)$, $B = \{B_1, B_2, \ldots, B_n\}$, $n = |B|$, let $x_i > 0$ denote the weight of subobject A_i $(1 \leq i \leq m)$, and let $y_i > 0$ denote the weight of subobject B_j $(1 \leq j \leq n)$. Assuming that the number of MOPs is p $(p \leq \min\{m, n\})$, denoted by $s_1, s_2, \ldots, s_p \in A \times B$, without loss of generality, and supposing they are $\langle A_1, B_1 \rangle$, $\langle A_2, B_2 \rangle, \ldots, \langle A_p, B_p \rangle$, with similarity degree $\mu_1, \mu_2, \ldots, \mu_p$, respectively, $\mu_i \geq \mu_0$ $(i = 1, \ldots, p)$, and the systematic similarity degree is $SS(O_A, O_B)$, then

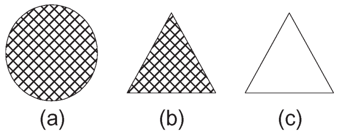


Fig. 4. Counterexample of triangle inequality.

$$SS(O_A, O_B) = \frac{\sum_{i=1}^{p} \mu_i x_i^2}{\sqrt{\sum_{i=1}^{m} x_i^2} \sqrt{\sum_{i=1}^{p} \mu_i^2 x_i^2 + \sum_{j=p+1}^{n} y_j^2}}. \quad (2)$$

We call function (2) as the "systematic similarity function." The function can be easily proved to meet the monotonicity conditions 1)–5) with the provision that $0 \leq \mu_0 \leq 0.5$.

The systematic similarity function (2) can be regarded as the cosine of the angle between the two N-dimension $N \geq m + n$ vectors $A = (x_1, x_2, x_3, \ldots, x_m, 0, \ldots, 0)_N$ and $B = (x_1\mu_1, x_2\mu_2, \ldots, x_p\mu_p, 0, \ldots, 0, y_{p+1}, y_{p+2}, \ldots, y_n)_N$. In terms of the Cauchy inequality, formula (2) happens to be in the range [0,1]. If there is no MOP with similarity degree greater than the threshold μ_0 $(0 < \mu_0 \leq 0.5)$ between two objects, formula (2) is equal to 0; if objects A and B are identical, their similarity degree is equal to 1. This property makes it possible to recursively compute the systematic similarity of the two objects through the systematic similarity of their subobjects.

It is easy to verify that formula (2) is not a commutative function, or, in other words, there exist some object A and object B that lead to $SS(A, B) \neq SS(B, A)$. In addition, the triangle inequality is also unsatisfied, i.e., there exist some objects A, B, C that cause $SS(A, B) + SS(B, C) \leq SS(A, C)$. Hence, the systematic similarity degree is not metric.

As has been discussed before, symmetry and triangle inequality are NOT necessary conditions that the systematic similarity function has to satisfy. When we assume that object A and object B are "alike," we usually choose one object as reference, which is called target object in this paper. As to the triangle inequality, Lin presents a counterillustration, as shown in Fig. 4 [36].

In Fig. 4, A and B are similar in their shades, B and C are similar in their shape, but A and C are not similar. In summary, it is not in contravention of reality and human experiences, although the systematic similarity function is not metric.

D. SSM and Other Similarity Models

The aim of the SSM is NOT to negate previous similarity models; instead, it presents a unified framework to cease the arguments around issues that are concerned with the characteristics of systematic similarity in the sense that it is, for example, capable of providing a satisfactory explanation for some natures of similarity such as asymmetries and invalidation of triangle inequality. In addition, the volatility of weight interprets the reason why similarity evaluation is sensitive to context, perspective, choice alternatives, and expertise.

Until now, the term "object" is default to discrete objects. For measuring the systematic similarity of continuous objects, for example, in the case of reflecting the fact that latex is more similar to milk than water, where structural alignment cannot be performed, it has no other choice but to select appropriate features and scale the distance in featural space, as the geometric model does, or compute according to common and differential features, as the contrast model does.

The most exciting characteristic of SSM is that it qualifies for comparing things that are richly structured. During the computation, it strictly follows the one-to-one mapping process and principle of systematicity, which has been proved by psychological tests and experiments [5], [37].

As has been mentioned before, the transformational model, which is a recent theoretical approach to similarity, provides a theoretical framework applicable to similarity judgments over representations of arbitrary form, including structured representations. According to this approach, the similarity between two entities is a function of the "complexity" required to "distort" or "transform" the representation of one into the representation of the other. The more simply the objects transform from one representation to the other, the more similar they are assumed to be. In many cases, SSM is able to draw the same conclusion on similarity estimation as the transformational model. However, this is not always the case. For the example given in [37], when elucidating the definition of "Representational Distortion," which is a fundamental concept of the transformational similarity model, the paper indicated that the transformational model will determine that XXXOOXO is more similar to OXOOXXX than to OXOOXX because OXOOXXX involves a mirror transformation of XXXOOXO, whereas OXOOXX involves a mirror transformation of XXXOOXO plus a deletion of the rightmost X from OXOOXXX. In contrast, within the framework of systematic similarity, supposing the weight of all the characters to be identical, SSM will determine that XXXOOXO is more similar to OXOOXX than to OXOOXXX [38]. The latent reason of this difference needs further study, but it has been reported that the transformational-based model cannot account for the experimental data that the structural alignment-based model fits well [37].

E. Systematic Similarity Computation Algorithm

A systematic similarity computing algorithm is presented in Fig. 9 to describe how to use the systematic similarity function [(2)] to compute the systematic similarity degree.

Before computation takes place, an assessment process is needed to ascertain the target object and source object from the given situations, or it may lead to a completely fault structural alignment process of the computation algorithm. For the example in Fig. 5, a prestructural alignment process should have chosen two smileys in object A and object B to be under similarity evaluation, or the algorithm will fail to perform structural matches from the starting point.

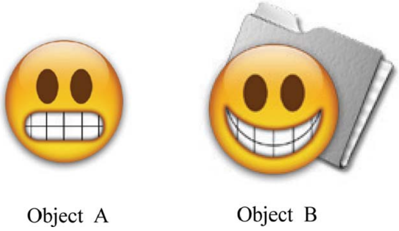


Fig. 5. Structural alignment example.

The algorithm is a recursive algorithm. The two inputs are semantic representations of object A and object B. The weights of all-level subobjects and the similarity degrees between entities have already been determined.

Line 1 to line 7 deals with the special input case, that is, at least one input parameter is entity. If so, then some functions are called to directly search or compute the systematic similarity, e.g., the EntitySimilarity function is called to search entity similarity degrees in knowledge base and return, the EntityObjectSimilarity function is called to compute the similarity degree between entities and a relation, and the ObjectEntitySimilarity function is called to compute the similarity degree between a relation and an entity.

Line 9 to line 26 is an essential part of the algorithm. It recursively constructs the systematic similarity matrix from line 11 to line 15, which is shown as the mapping process in Fig. 1. The systematic similarity matrix not only enables the computation process to follow the human's preference in similarity match by a recursive mechanism but also serves as one of the determinant factors for constructing MOPs. It entirely reveals the latent consistency between structure mapping theory and principle of systematicity. From line 16 to line 20, the algorithm finds all the MOPs of the two objects with regard of similarity matrix and semantic relation. Because of the variety of semantic relations, this is the most important and complicated step through which to construct the matched vectors of the two objects in the algorithm. Line 21 deals with a special case that both objects are unary relations. From line 22 to line 23, unmatched vectors are constructed. After that, formula (2) is adopted to compute the systematic similarity between two objects and return the result. Line 16 to line 25 imitates the evaluation stage in Fig. 1.

To fully understand the systematic similarity computing algorithm, let us see what happens on our sentence match examples. Before the algorithm is called, both target sentence and source sentence should have been semantically parsed, that is, after a series of syntactic and semantic analysis, both target sentence and source sentence have been transformed into the object representation conventions of this paper, which may appear as follows (see also Figs. 6 and 7):

Target sentence: TARGET_SENTENCE(IF2(THOUGH (AND(BESTOW(I, BESTOW, GOOD(all, my), FEED(FEED, POOR)), GIVE(I, GIVE, BODY(my), BE_BURNED))), IF1(HAVE_NOT, CHARITY)), THEN(PROFITETH(IT, PROFITETH, ME, NOTHING)))

Further assuming that the weights of the entities are as follows, these values are simply set according to the authors' preference (Table III).

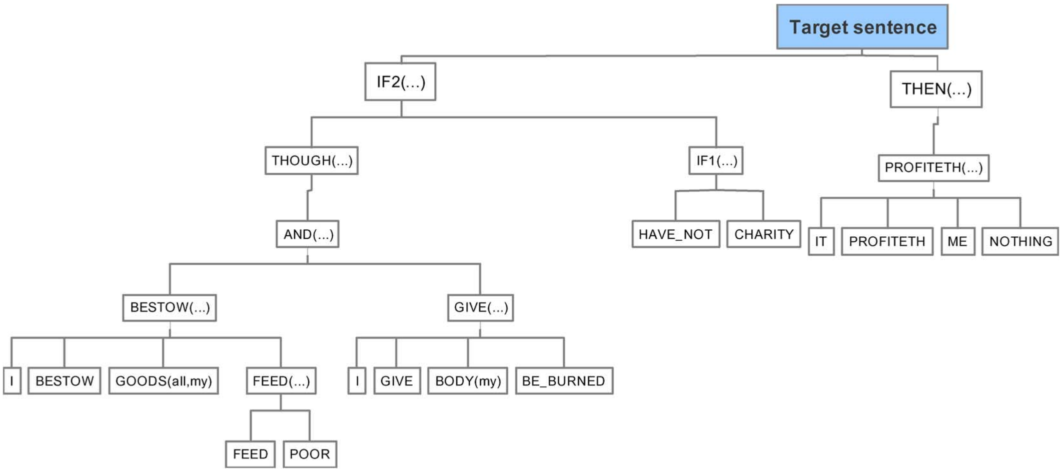
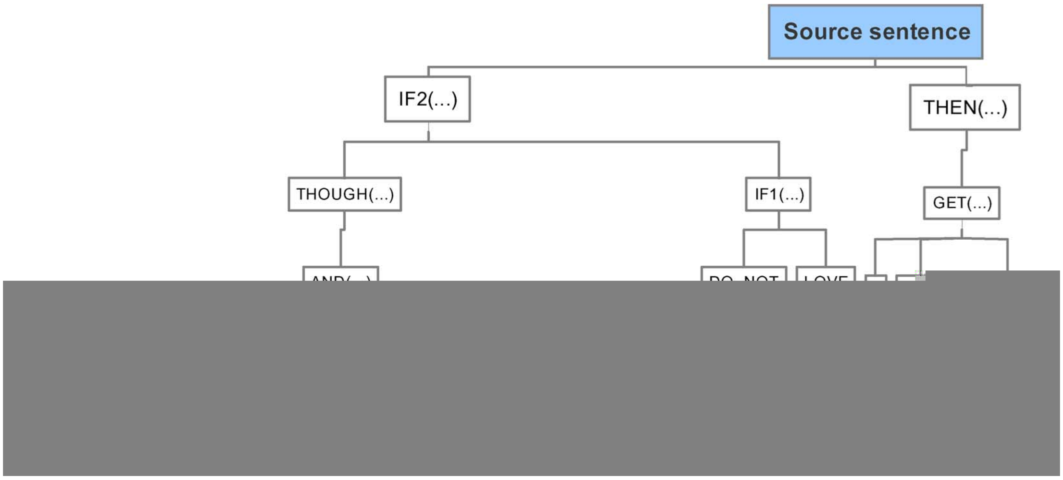


Fig. 6. Target sentence.



[redacted]

Fig. 7. Source sentence.

TABLE III
WEIGHTS OF ENTITIES OF THE SENTENCE MATCH EXAMPLE

Target sentence		Source sentence	
Entity	**Weight**	**Entity**	**Weight**
I	6.0	I	6.0
BESTOW	9.0	GIVE	7.0
GOODS	6.0	THING	5.0
FEED	8.0	POOR	7.0
POOR	7.0	GO_TO	7.0
GIVE	7.0	STAKE	7.0
BODY	8.0	BE_BURNED	9.0
BE_BURNED	9.0	DO_NOT	6.0
HAVE_NOT	5.0	LOVE	8.0
CHARITY	9.0	GOT	7.0
IT	4.0	NOWHERE	2.0
PROFITETH	8.0		
ME	3.0		
NOTHING	1.0		

Source sentence: SOURCE_SENTENCE(IF2(THOUGH (AND(GIVE(I, GIVE, THING(all, my), POOR), GO_TO(I, GO_TO, STAKE, BE_BURNED(martyr)))), IF1(DO_NOT, LOVE)), THEN(GET(I, GOT, NOWHERE)))

The MOPs and their associated similarity degrees between entities may look as follows (μ_0 is assumed equal to 0.5):

$\langle \text{I}, \text{I} \rangle$, 1.0;
$\langle \text{BESTOW}, \text{GIVE} \rangle 0.95$;
$\langle \text{GOODS(all, my)}, \text{THING(all, my)} \rangle 0.9$;
$\langle \text{POOR}, \text{POOR} \rangle 1.0$;
$\langle \text{BE\_BURNED}, \text{BE\_BURNED(martyr)} \rangle 1.0$;
$\langle \text{HAVE\_NOT}, \text{DO\_NOT} \rangle 0.8$;
$\langle \text{CHARITY}, \text{LOVE} \rangle 0.6$;
$\langle \text{I}, \text{ME} \rangle 0.9$;
$\langle \text{I}, \text{IT} \rangle 0.5$;
$\langle \text{PROFITETH}, \text{GOT} \rangle 0.6$;
$\langle \text{NOTHING}, \text{NOWHERE} \rangle 0.55$.

In the case of an entity with a relation, their systematic similarity may be measured as

$$Similarity\,(\text{POOR}, \text{FEED}(\text{FEED}, \text{POOR})) = \frac{\mu^*_{\langle \text{POOR},\text{POOR} \rangle}\, weight_{\text{POOR}}}{\sqrt{weight^2_{\text{FEED}} + weight^2_{\text{POOR}}}}. \quad (3)$$

Using formula (3), we get

$$\langle \text{POOR}, \text{FEED}(\text{FEED}, \text{POOR})\rangle 0.66.$$

This method is easy to be generalized for the case of an entity with a more complex relation. In that case, an entity match process must firstly be carried out by which to decide their similarity degree [for example, $\mu_{\langle \text{POOR,POOR}\rangle}$in (3)]. This value is then served as a starting value of a bottom-up induction of the final systematic similarity value with the function like formula (3). This process can be seen as a special case of the systematic similarity computation algorithm and has been implemented in the functions EntityObjectSimilarity and ObjectEntitySimilarity.

Assuming that in the semantic tree the weight of an object is assigned to be the largest weight of its subobjects, then the algorithm sequentially computes the systematic similarity degree between relations as follows:

$\langle \text{BESTOW}(\ldots), \text{GIVE}(\ldots)\rangle 0.99$;
$\langle \text{GIVE}(\ldots), \text{GO\_TO}(\ldots)\rangle 0.53$;
$\langle \text{AND}(\ldots), \text{AND}(\ldots)\rangle 0.96$;
$\langle \text{THOUGH}(\ldots), \text{THOUGH}(\ldots)\rangle 0.96$ (by line 21 in the algorithm);
$\langle \text{IF1}(\ldots), \text{IF1}(\ldots)\rangle 0.99$;
$\langle \text{IF2}(\ldots), \text{IF2}(\ldots)\rangle 0.9999$;
$\langle \text{PROFITETH}(\ldots), \text{GET}(\ldots)\rangle 0.837$;
$\langle \text{THEN}(\ldots), \text{THEN}(\ldots)\rangle 0.837$ (by line 21 in the algorithm);
$\langle \text{TARGET SENTENCE}, \text{SOURCE SENTENCE}\rangle 0.996223$.

The result tells us that although these two sentences look quite different on the lexical surface, they almost have the same meaning (or have very high semantic similarity degree).

There are many problems left in the sentence match example, for instance, the semantic representations of sentences are very informal, some binary relations are arbitrarily changed into unary relations for the convenience of describing how the algorithm deals with that case, etc. Anyway, this example clearly shows that the systematic similarity algorithm is nothing but an imitation of the human's similarity evaluation process. It simplified the SME by means of an algebraic function and an associated recursive mechanism. In addition, with the recursive mechanism, it strictly follows the structural mapping theory and performs the one-to-one mapping process on every hierarchy of the systematic structures.

F. Universality of SSM

The authors believe that SSM is a domain-independent framework for the following reasons: the theoretical foundation of the model (structure mapping theory and the principle of systematicity), which have been successfully applied to many areas such as categorization, decision making, artificial intelligence, etc., and widely accepted as basic laws for the assessment of either conceptual similarity or perceptual similarity [25], are served as a fundamental description of the nature of the human's cognitive ability, which is independent of cognitive objectives, whether it is text, image, or video. Moreover, the systematic similarity computation algorithm is simply an imitative process of human similarity evaluation. In a word, SSM is a model of the human cognitive process. The object representation convention of the model is not restricted to any particular specified content. These factors allow SSM to act as a domain-general model.

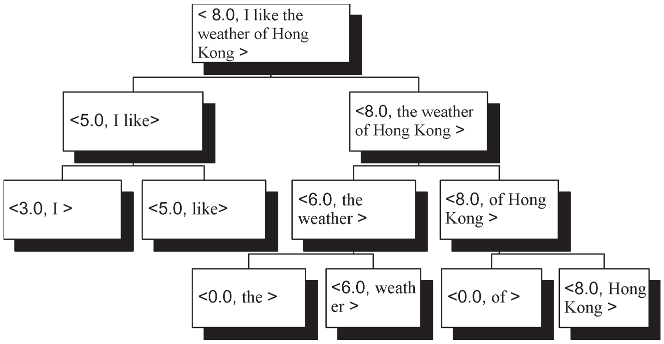


Fig. 8. Query (or document) representation of SSM for IR. In this example, the weights of father nodes are set by the maximum weight of its child nodes.

Although the model has not been applied in image and video retrieval, we have performed a series of preliminary experiments on text retrieval.

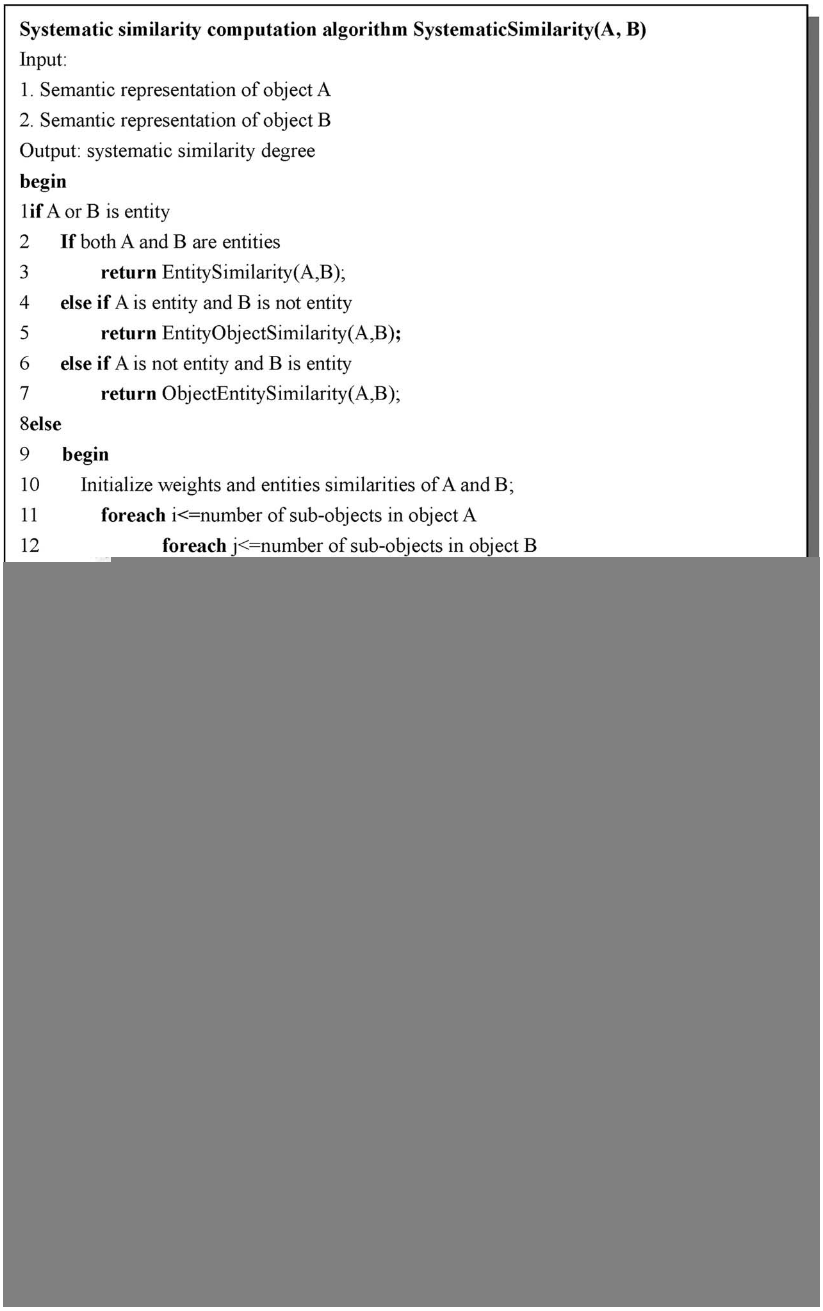
V. Case Study: SSM for Text Retrieval

In this section, we will illustrate the typical applications of SSM by its usage as an information retrieval model.

Nowadays, with the dramatic growth of electronic data, more and more applications adopt example-based methods in which a similarity measurement function acts as the core of the system. Such applications include text retrieval, image retrieval, example-based machine translation, etc. Since SSM is a very generic framework, a systematic similarity function [(2)] is still needed to combine all the factors, among which systematic structure, entity similarity, and weight have to be redefined and retrieved from domain-specific features in adopting the SSM-based model in different applications of the specific domain.

An information retrieval model is a mathematical framework that defines and explains major phases of the information retrieval process, including document representation, user query representation, ranking of retrieved documents, etc. Formally, it is defined as a quadruple $[D, Q, F, R(q_i, d_j)]$, where D is a set composed of representations for documents, Q is a set composed of representations for user queries, F is a framework for modeling document representations, queries, and their relationships, and $R(q_i, d_j)$ is a ranking function that associates a number $\in \mathbf{R}$ with a query and a document.

Since a precise semantic parsing of web text is still a remote dream for information retrieval, we adopted syntactic parsing to approximate semantic parsing. In SSM for information retrieval, both user query and document are represented by a binary tuple $\langle w, \{c_1, c_2, \ldots, c_n\}\rangle$, where w is the overall weight of the total subcomponents $c_1, c_2, \ldots, c_n$, and $\{c_1, c_2, \ldots, c_n\}$ is an ordered set of n constituting components. Each c_i $(1 \leq i \leq n)$ is a component that also takes the form of a binary tuple $\langle w_i, \{c_{i1}, c_{i2}, \ldots, c_{in}\}\rangle$. A component is an entity if n is equal to 1, i.e., $\langle w, \{c_1\}\rangle$, where c_1 is a term and w is a term weighting in IR. For example, the text snippet "I like the weather of



```
Systematic similarity computation algorithm SystematicSimilarity(A, B)
Input:
1. Semantic representation of object A
2. Semantic representation of object B
Output: systematic similarity degree
begin
1if A or B is entity
2     If both A and B are entities
3           return EntitySimilarity(A,B);
4     else if A is entity and B is not entity
5           return EntityObjectSimilarity(A,B);
6     else if A is not entity and B is entity
7           return ObjectEntitySimilarity(A,B);
8else
9     begin
10       Initialize weights and entities similarities of A and B;
11          foreach i<=number of sub-objects in object A
12                   foreach j<=number of sub-objects in object B
```

Fig. 9. Systematic similarity computation algorithm.

Hong Kong" is represented by the nested binary tuple shown in Fig. 8.

To compute the relevance score between user query and document, SSM uses the systematic similarity function [(2)]. The systematic similarity computing algorithm is very similar to the algorithm presented in Fig. 9. Since the semantic representations of queries and web documents have been replaced by syntactic representations, this brings about some differences in the implementation of the algorithm. 1) The algorithm takes three input parameters, i.e., thesaurus and set representation of sentence A and sentence B. A thesaurus is a global system resource that contains all words with importance and similarity degree between two words. 2) Since both query and document are not semantically parsed, the MOPs have to be decided by their similarity degrees and syntactic roles. In fact, in the implemented algorithm, line 17 is replaced as follows. For each subobject $a_i \in A$, let $b_j = \arg\max_{b_j \in B}(similarity(a_i, b_j))$. If $similarity(a_i, b_j)$ is greater than some predefined threshold $0 < \mu_0 < 1$, then $\langle a_i, b_j \rangle$ forms a MOP, that is, the syntactic role constrains are ignored.

Another interesting attribute of SSM is that it can be considered as a generic case of the vector space model (VSM) [39], which has been widely applied in IR fields over years. In VSM, a document is represented by a set of index terms with weights (vector of terms), and so is user query; a weight expresses the relative importance of the term with respect to the document. Letting $\vec{A}$ denote the user query vector $(x_1, x_2, \ldots, x_N)$ and

TABLE IV
QUERY VECTOR AND DOCUMENT VECTOR IN VSM

	t_1	t_2	t_3	t_4	…	t_{N-1}	t_N
q	x_1	x_2	0	x_4	0	0	0
d	x_1	0	x_3	0	0	x_{N-1}	x_N

TABLE V
QUERY VECTOR AND DOCUMENT VECTOR IN SSM

	Terms in MSCPs			Terms not in MSCPs		
	t_2	t_1	t_3	t_4	t_{N-1}	t_N
Q	x_2	x_1	0	x_4	0	0
D	$\mu \cdot x_2$	$1 \cdot x_1$	0	0	x_{N-1}	x_N

$\vec{B}$ denote the document vector $(y_1, y_2, \ldots, y_N)$, their relevance score (similarity) is computed by

$$Similarity(\vec{A}, \vec{B}) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{4}$$

i.e., the cosine of the angle between the two vectors. Normally, the query vector and the document vector are represented by N-dimension vectors. For example, let $t_1, t_2, \ldots, t_N$ denote all the terms of an IR system vocabulary. Their weights are, respectively, denoted by $x_1, x_2, \ldots, x_N$. A user query q is $t_2 t_1 t_4$. A document d is $t_1 t_3 t_{N-1} t_N$. These vectors are shown in Table IV.

In comparison, under the schema of SSM, for the same q and d, to simplify our discussion, assuming q is syntactically parsed to be $\{\{t_2\}, \{t_1\}, \{t_4\}\}$, and d is $\{\{t_1\}, \{t_3\}, \{t_{N-1}\}, \{t_N\}\}$, and further assuming that $similarity(t_2, t_3) = \mu (\mu \geq 0.5)$ (and $similarity(t_1, t_1) = 1$), or in other words, for systems q and d, $\langle t_1, t_1 \rangle$, $\langle t_2, t_3 \rangle$ are MOPs, then vectors q and d are constructed in the following way instead, as shown in Table V.

The vectors in Table V can be considered as a transformation from that in Table IV. Instead of being represented by N-dimension vectors, both q and d are represented by M-dimension $(M = |q \cup d|)$ vectors. For vector q, each dimension is set to be the importance of the component. For vector d, a group of new parameters, i.e., μ, which not only allows exact term matching (for t_1) but also enables similar term matching (for t_2 in q and t_3 in d), is introduced into the model, as fulfilled by removing (set to zero) x_3 from the unmatched dimension (t_3-dimension) and putting $\mu \cdot x_2$ in a similar matched dimension (t_2-dimension), and so is t_1. The idea behind this is that the similarity degree of two elements can be viewed as an extent to which they can be replaced with each other in some context.

Except for differences in dimension definitions and constructing way for vectors, both VSM and SSM compute the cosine of the angle of two vectors. In fact, an essential difference exists in the range of μ. If μ is restricted to either 1 or 0, and the terms are presumed to be independent of one another, then VSM and SSM are equivalent. By contrast, SSM permits μ to be any real number in the range $[\mu_0, 1]$, which supplies VSM with structural semantic similar matching. Therefore, VSM can be regarded as a special case of SSM.

VSM is often criticized for lack of theoretical soundness [40]. The previous discussion reveals that it is actually a special case of SSM, which is deduced from the computation of the systematic similarity degree between two objects. In addition, the terms are no longer independent of one another, but demonstrate dependency by hierarchical structures. The order of terms is reserved, which can be taken into consideration by more accurate relevance computation. As to the problem of polysemy and synonymy, SSM also has the capability of providing a novel solution. In IR, the problem of polysemy, also known as word sense ambiguity, often leads to a wrong response of user query. However, in the framework of SSM, if two terms (a term in a query and the other in a document) have the same word sense with different lexical form, they could achieve some semantic similarity degree (μ) by some semantic similarity measurement method, which would accordingly increase the overall relevance score between user query and document. For example, for the query "how about the weather of Hong Kong," the text snippet "The climate of Hong Kong is typical tropical" would achieve a higher relevance score because of the entity MOPs $\langle \text{Hong Kong}, \text{Hong Kong} \rangle$, $\langle \text{weather}, \text{climate} \rangle$, and consequent object MOP $\langle$the weather of Hong Kong, the climate of Hong Kong$\rangle$. For terms with the same lexical forms, noticing the fact that a sense of a word is determined by its context, although their semantic similarity is one, their father object's semantic similarity could be lowered down by a different context (or become higher by similar context). For example, for phrases "red alert" and "red apple," although both phrases contain the word "red," and $\mu(red, red) = 1$, the similarity degree of them will be decreased by more important words ("alert" and "apple") that are not in MOP so that these two phrases cannot become MOP.

VI. EXPERIMENTS AND DISCUSSION

As a new measurement for a psychological concept, SSM should have been directly tested by psychological experiments. However, due to the same reason, such experiments are very difficult to be performed when a precise mathematical model is imposed on such a fuzzy and abstract concept as similarity. So we indirectly test the SSM by its performance as an information retrieval model in three passage ranking experiments.

A. Experiment 1: Simple Query Matching Experiment

In the first experiment, we compare the top ten precision of the SSM with VSM by a simple query matching experiment.

Forty-five undergraduate students (divided into two groups) participated in this study. The first group consists of 40 subjects whose duty is to collect 20 000 natural language questions on health care with reference to a preappointed list of health care websites. The other five subjects are required to ask 1500 natural language questions with reference to the same websites as the first group.

A thesaurus that stores all words in HowNet [41] with semantic similar words and their similar degrees is constructed [42]. The term weight is quantified by tf*idf measurement. A query matching algorithm, which is based on the algorithm shown in Fig. 9, is developed to compute the semantic similarity degree between user query and database query with the aid of the semantic similarity between words. However, both user query and database query are not parsed.

One hundred randomly collected real user queries are used to compare the SSM-based ranking algorithm and the VSM-based ranking algorithm. (Because VSM is a special case of SSM, such a comparison is easy to execute.) We put each query into the query matching system, and it is decided by the human whether the correlative database query is ranked in the top ten. The experimental result shows that the precision of the VSM-based algorithm is 52%, and while the SSM-based algorithm increases to 61%, a 9% enhancement is achieved [42].

B. Experiment 2: Experiments on SSM for Information Retrieval

In the intelligent information retrieval system for tour domain, SSM is adopted for text representation and relevance score computation [43]. The semantic similarity degree between Chinese words has been computed based on Hownet [42]; 101 500 Chinese tourism documents have been crawled, segmented, Part-of-speech tagged, named entity tagged, and syntactically parsed, and all text snippets are indexed in the text database. A preliminary evaluation has been undertaken by 2000 user queries, which are guaranteed to have answers in these documents. The system achieved 45% top ten precision.

C. Experiment 3: Experiments in TREC 2005 QA Track

At TREC 2005, a question answering system in which SSM is adopted as the core component of sentence similarity calculation (Insun05QA), which submitted answers to three types of questions (i.e., factoid questions, list questions, and others questions), participated in the Main Task [44].

In the system, GATE is used to accomplish name entity reorganization in the preprocess stage of questions and documents [45]. Minipar is adopted for question analysis and also used to parse the relevant passage in the answer extraction module [46]. The document retrieval module is developed based on SMART [47]. Semantic similarities between English words are computed by an information-content-based method based on Wordnet [48], [56]. Term weights are obtained by the method of a variation of standard tf*idf scheme.

The evaluation result is shown in Table VI.

TABLE VI
PERFORMANCE OF Insun05QA IN TREC 2005

		Insun2005QA1
Average per-series score		0.187
Factoid questions	Number of correct	106
	Number of unsupported	15
	Number of inexact	16
	Number of wrong	225
	Accuracy	0.293 (median accuracy scores 0.153)
	Precision of NIL	0.057
[redacted]	[redacted]	[redacted]

In 71 participants, the system ranks fifth for factoid questions, seventh for list questions, and eighth in average per-series score [49].

D. Discussion

In experiment 1, because both user query and database query adopt the same term weighting method and are not parsed, the improvement of SSM over VSM mainly comes from the introduction of the semantic similarity of SSM. However, in experiments 2 and 3, SSM only achieves a comparable performance to other models and does not show a compelling improvement. The reason for this comes from the fact that not only are there too many processing steps in the question answering system that may magnify the probability of faulty retrieval but also one important theoretical requirement of SSM is not met in the implemented question answering systems, i.e., in SSM, *both objects need to be semantically parsed.* Until now, what we have is only a syntactic parser that is far from perfect. There is still a long way for SSM to give its full play.

VII. RELATED WORKS

Many works have focused on calculating the similarity of data objects. The approaches that use a single type of relationship to calculate the similarity of document-query objects include VSM [39], generalized VSM [50], [51], latent semantic indexing [52], query expansion [53], and dynamic vector space modification [54]. These can be viewed as variations of a general algorithm that projects documents and queries into a vector space using singular vectors. They differ in how the vectors are constructed and how weights are assigned. Dice, Jaccard, Cosine [55] and information entropy [56] measurements are a few classical methods to measure the similarity.

The research works that only use a single type of relationship to measure the similarity of data objects may run into serious problems when various information applications require a more real and accurate similarity measuring method. It should be a better choice to handle the multiple types of data objects and their relationship in an integrated manner. Das *et al.* define the similarity between attributes in large data sets and adopt two basic approaches for attribute similarity, i.e., internal and external measures [57]. Jeh and Widom measure the similarity of the structural context in which objects occur based on their relationships with other objects [58]. Gavesan proposes a similarity measure by adding a hierarchy describing the relationships among domain elements. The "semantic knowledge" in the hierarchy helps to identify objects sharing common characteristics, leading to the improved measures of similarity [59]. Except modeling the semantic similarity within a single ontology, the work in [60] presents an approach to computing the semantic similarity that relaxes the requirement of a single ontology and attempts to compute the semantic similarity among entity classes from different ontologies. In addition, as many organizations are confronted with the challenge of interoperating among multiple independent database systems, the primary issue is not to efficiently process the relevant data but to determine which data are relevant. The research presented in [61] works for identifying objects in different databases that are semantically related, and then resolving the schematic differences among these objects.

Recently, the approaches that tried to calculate the similarity of two data objects by measuring the similarity of their related data objects are also focused. For example, Raghavan and Sever tried to measure the similarity of two queries by calculating the similarity relationship of their corresponding search lists [62]. Beeferman and Berger clustered queries using the similarity of their clicked web pages and clustered web pages using the similarity of the queries that lead to the selection of the web pages [63]. Davison had proposed another much related idea [64]. In his two-page short paper, he analyzed multiple term-document relationships by expanding the traditional document-term matrix into a matrix with term–term, doc–doc, term–doc, and doc–term submatrices. He proposed that the links of the search objects (web page or terms) in the expanded matrix could be emphasized. With enough emphasis, the principal eigenvector of the extended matrix will have the search object on top with the remaining objects ordered according to their relevance to the search object, combining intertype relationship over the intratype and intertype relationship matrix.

VIII. Conclusion

In this paper, a new theoretical framework on the measurement of systematic similarity has been presented and discussed. SSM, which is mainly inspired by the contrast model of similarity and structure mapping theory in psychology, not only takes a simple and concise form but also enjoys compatibility to current similarity models, and above all, it is a universal framework that can be applied to any example-based systems. The new model settles the controversy on similarity measurement by the following affirmations.

1) Systematic similarity is NOT a metric measurement.
2) The systematic similarity degree between complex objects that are identical to each other is the same as that of simpler identical objects.
3) The systematic similarity degree of system A and system B is a scalar value in [0,1], and 1 is achieved iff A is a copy of B or vice versa.
4) The similarity degree of compared objects increases with its similar subobjects and decreases with its different subobjects.
5) The systematic similarity degree takes the identical form as (2); the differences of similarity measurement for different domains only exist in weight and entity similarity measurement, which can be regarded as functions of domain-specific features.
6) When the systematic similarity degree is estimated, the weight can be varied with different relations and structures.

Although some of the above affirmations have not been proved by psychological experiments, our tentative attempts on information retrieval modeling have shown the validity of the new model. Research on the measurement of systematic similarity is ongoing, and many theoretical problems deserve academic efforts, for example, the uniqueness of the systematic similarity degree formula, the weight variance modes, etc. These problems are awaiting psychologists, linguists, and computer scientists for further cooperation.

Acknowledgment

The authors would like to thank J. Nie, X. Wang, X. Sui, and X. Dong for the suggestions and comments on a preliminary draft of this paper; the anonymous reviewers for their insightful comments and suggestions; to the people who contribute to system integration and testing (P. Fung, Y. Yang, P. Cheung, and X. Wang for help on simple query matching experiments); the Web information processing team of Research Center of Language Technologies in School of Computer Science and Technology (RCLT), Harbin Institute of Technology, who have accomplished the integration and testing of the intelligent information retrieval system for tour domain (C. Sun, G. Sun, P. Li, W. Jiang, J. Zhao, K. Sun, Q. Dong, M. Li, X. Du, W. Jia, X. Lv, X. Jia, Y. Pu, X. Sun, Y. Mao, B. Shao, and G. Bao); the TREC QA team of RCLT (P. Li, Y. Zhao, and B. Wang); and the professors of RCLT (Z. Xu, B. Liu, L. Lin, and Y. Liu).

References

[1] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Past, present, and future," *J. Vis. Commun. Image Represent.*, vol. 10, pp. 1–23, 1999.

[2] S.-C. S. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 59–74, Jan. 2003.

[3] J. Gu, L. Lu, R. Cai, H.-J. Zhang, and J. Yang, "Dominant feature vectors based audio similarity measure," in *Proc. PCM*, Tokyo, Japan, Nov. 30–Dec. 3, 2004, pp. 890–897.

[4] T. Mandl, "Learning similarity functions in information retrieval," in *Proc. EUFIT 6th Eur. Congr. Intell. Tech. Soft Comput.*, H. J. Zimmermann, Ed., Aachen, Germany, 1998, pp. 771–775.

[5] D. Gentner, "The mechanisms of analogical learning," in *Similarity and Analogical Reasoning*, S. Vosniadou and A. Ortony, Eds. London, U.K.: Cambridge Univ. Press, 1989.

[6] L. B. Smith, "From global similarities to kinds of similarities: The construction of dimensions in development," in *Similarity and Analogy*, S. Vosniadou and A. Ortony, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 146–178.

[7] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.

[8] W. S. Torgerson, "Multidimensional scaling of similarity," *Psychometrika*, vol. 30, pp. 379–393, 1965.

[9] R. L. Goldstone, D. L. Medin, and D. Gentner, "Relational similarity and the nonindependence of features in similarity judgments," *Cogn. Psychol.*, vol. 23, no. 2, pp. 222–262, Apr. 1991.

[10] D. Gentner and A. B. Markman, "Structure mapping in analogy and similarity," *Amer. Psychol.*, vol. 52, no. 1, pp. 45–56, Jan. 1997.

[11] R. L. Goldstone, "Similarity," in *MIT Encyclopedia of the Cognitive Sciences*, R. A. Wilson and F. C. Keil, Eds. Cambridge, MA: MIT Press.

[12] R. M. Nosofsky, "Similarity scaling and cognitive process models," *Annu. Rev. Psychol.*, vol. 43, pp. 25–53, 1992.

[13] R. M. Nosofsky, "Stimulus bias, asymmetric similarity, and classification," *Cogn. Psychol.*, vol. 23, no. 1, pp. 94–140, Jan. 1991.

[14] D. H. Krantz and A. Tversky, "Similarity of rectangles: An analysis of subjective dimensions," *J. Math. Psychol.*, vol. 12, no. 1, pp. 4–34, Feb. 1975.

[15] A. Tversky and I. Gati, "Studies of similarity," in *Cognition and Categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1978, pp. 79–98.

[16] R. L. Goldstone, "Similarity, interactive activation, and mapping," *J. Exper. Psychol., Learn., Mem., Cogn.*, vol. 20, no. 1, pp. 3–28, Jan. 1994.

[17] A. B. Markman and D. Gentner, "Structural alignment during similarity comparisons," *Cogn. Psychol.*, vol. 25, no. 4, pp. 431–467, Oct. 1993.

[18] R. Yang, P. Kalnis, and A. K. H. Tung, "Similarity evaluation on tree-structured data," in *Proc. ACM SIGMOD Conf.*, Baltimore, MD, 2005, pp. 754–765.

[19] T. Wang, Y. Rui, S. M. Hu, and J. G. Sun, "Adaptive tree similarity learning for image retrieval," *ACM Multimedia Syst. J. (MMSJ)*, vol. 9, no. 2, pp. 131–143, Aug. 2003.

[20] U. Hahn and N. Chater, "Concepts and similarity," in *Knowledge, Concepts and Categories*, K. Lamberts and D. Shanks, Eds. Hove, U.K.: Psychology Press, 1997, pp. 43–92.

[21] U. Hahn, N. Chater, and L. B. C. Richardson, "Similarity as transformation," *Cognition*, vol. 87, no. 1, pp. 1–32, Feb. 2003.

[22] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, Sep. 1999.

[23] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cogn. Sci.*, vol. 7, no. 2, pp. 155–170, Apr.–Jun. 1983.

[24] R. M. French, "The computational modeling of analogy-making," *Trends Cogn. Sci.*, vol. 6, no. 5, pp. 200–205, May 2002.

[25] D. Gentner and A. B. Markman, "Similarity is like analogy: Structural alignment in comparison," in *Similarity in Language, Thought and Perception*, C. Cacciari, Ed. Brussels, Belgium: BREPOLS, pp. 111–147.

[26] R. Schumacher and D. Gentner, "Similarity-based remindings: The effects of similarity and interitem difference," presented at the Annu. Meeting Midwestern Psychological Association, Chicago, IL, 1987.

[27] M. J. Rattermann and D. Gentner, "Analogy and similarity: Determinants of accessibility and inferential soundness," in *Proc. 9th Annu. Conf. Cogn. Sci. Soc.*, Seattle, WA, 1987, pp. 23–25.

[28] B. Falkenhainer, K. D. Forbus, and D. Gentner, "The structure-mapping engine: Algorithm and examples," *Artif. Intell.*, vol. 41, no. 1, pp. 1–63, Nov. 1990.

[29] B. Falkenhainer, "Structure mapping engine implementation," *SME Implementation*, 2005. [Online]. Available: http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/reasonng/analogy/sme/0.html

[30] M. D. Lee, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text document similarity," in *Proc. 27th Annu. Meeting CogSci*, B. G. Bara, L. Barsalou, and M. Bucciarelli, Eds., 2005, pp. 1254–1259.

[31] A. Bifet, C. Castillo, P. A. Chirita, and I. Weber, "An analysis of factors used in a search engine's ranking," presented at the 1st Int. Workshop Adversarial Information Retrieval Web, Chiba, Japan, 2005.

[32] Wikipedia Contributors, *The Message (Bible)*. Wikipedia, The free encyclopedia, accessed October 8, 2006. [Online]. Available: http://en.wikipedia.org/w/index.php?title=The_Message_%28Bible%28Bible%29&oldid=75917151

[33] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 17–30, Jan./Feb. 1989.

[34] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, Jul./Aug. 2003.

[35] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[36] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.

[37] L. B. Larkey and A. B. Markman, "Processes of similarity judgment," *Cogn. Sci.*, vol. 29, no. 6, pp. 55–70, 2005.

[38] S. Imai, "Pattern similarity and cognitive transformations," *Acta Psychol.*, vol. 41, no. 6, pp. 433–447, Oct. 1977.

[39] G. Salton, *Automatic Information Organization and Retrieval*. New York: McGraw-Hill, 1968.

[40] F. Crestani, M. Lalmas, C. J. V. Rijsbergen, and I. Campbell, "Is this document relevant?...Probably: A survey of probabilistic models in information retrieval," *ACM Comput. Surv.*, vol. 30, no. 4, pp. 528–552, Dec. 1998.

[41] Z. D. Dong and Q. Dong, *Hownet*. [Online]. Available: http://www.keenage.com/

[42] Y. Guan, X. Wang, X. Kong, and J. Zhao, "Quantifying semantic similarity of Chinese words from HowNet," in *Proc. IEEE ICMLC*, Beijing, China, 2002, vol. 1, pp. 234–239.

[43] C. Sun, Y. Guan, X. Wang, Q. Wang, and T. Liu, "InsunTourQA: A restricted-domain question answering system," *J. Comput. Inf. Syst.*, vol. 3, no. 4, pp. 1581–1590, 2007.

[44] Y. Zhao, Z. Xu, Y. Guan, and P. Li, "Insun05QA on QA track of TREC2005," in *Proc. TREC*, 2005. [Online]. Available: http://trec.nist.gov/pubs/trec14/papers/harbin-it.qa.pdf

[45] H. Cunningham, "GATE, a general architecture for text engineering," *Comput. Humanit.*, vol. 36, no. 2, pp. 223–254, May 2002.

[46] D. Lin, "A dependency-based method for evaluating broad-coverage parsers," in *Proc. IJCAI*, 1995, pp. 1420–1427.

[47] C. Buckley, A. Singhal, and M. Mitra, "Using query zoning and correlation with SMART: TREC-5," in *Proc. 5th Text Retrieval Conf.*, 1996, pp. 105–118.

[48] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[49] E. M. Voorhees and H. T. Dang, "Overview of the TREC 2005 question answering track," in *Proc. TREC*, 2005. [Online]. Available: http://trec.nist.gov/pubs/trec14/papers/QA.OVERVIEW.pdf

[50] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong, "On modeling of information retrieval concepts in vector space," *ACM Trans. Database Syst.*, vol. 12, no. 2, pp. 299–321, 1987.

[51] A. Bernstein, E. Kaufmann, C. Burki, and M. Klein, "How similar is it? Towards personalized similarity measures in ontologies," in *Proc. 7 Internationale Tagung Wirtschaftsinformatik*, Feb. 2005, pp. 1347–1366.

[52] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve information retrieval," in *Proc. Conf. Human Factors Comput. Syst.*, Washington, DC, May 1988, pp. 281–285.

[53] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System—Experiments in Automatic Document Processing*, G. Salton, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[54] T. L. Brauen, "Document vector modification," in *The Smart Retrieval System-Experiments in Automatic Document Processing*, G. Salton, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971, ch. 24.

[55] E. Rasmussen, "Clustering algorithms," in *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[56] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. IJCAI*, 1995, pp. 448–453.

[57] G. Das, H. Mannila, and P. Ronkainen, "Similarity of attributes by external probes," in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining*, 1998, pp. 23–29.

[58] J. Jeh and J. Widom, "SimRank: A measure of structural context similarity," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Edmonton, AB, Canada, Jul. 23–26, 2002, pp. 538–543.

[59] P. Gavesan, H. Garcia-Molina, and J. Widom, "Exploiting hierarchical domain structure to compute similarity," *ACM Trans. Inf. Syst.*, vol. 21, no. 1, pp. 64–93, Jan. 2003.

[60] M. A. Rodriguez and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 442–456, Mar./Apr. 2003.

[61] V. Kashyap and A. Sheth, "Semantic and schematic similarities between database objects: A context-based approach," *VLDB J.*, vol. 5, no. 4, pp. 276–304, 1996.

[62] A. Popescul, G. Flake, S. Lawrence, L. H. Ungar, and C. L. Giles, "Clustering and identifying temporal trends in document database," in *Proc. IEEE Advances Digital Libraries*, Washington, DC, 2000, pp. 173–182.

[63] D. Beefermand and A. Berger, “Agglomerative clustering of a search engine query log,” in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Boston, MA, 2000, pp. 407–415.

[64] B. D. Davison, “Toward a unification of text and link analysis,” in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Toronto, ON, Canada, 2003, pp. 367–368.

Yi Guan (M’05) received the B.Sc. degree in computer science and technology from Tianjin University, Tianjin, China, in 1992, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1999.

In 1996, he was an Invited Visiting Scholar with Canotec Co., Japan. In 2000, he was a Research Associate with the Human Language Technology Center, Hong Kong University of Science and Technology, Kowloon. In 2001, he was a Research Scientist with Weniwen.com, Hong Kong. In October 2001, he was an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. Since October 2006, he has been a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include question answering, statistical language processing, parsing, and text mining.

Xiaolong Wang received the B.E. degree in computer science from the Harbin Institute of Electrical Technology, Harbin, China, in 1982, the M.E. degree in computer architecture from Tianjin University, Tianjin, China, in 1984, and the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology in 1989.

He was an Assistant Lecturer in 1984 and an Associate Professor in 1990 with the Harbin Institute of Technology. From 1998 to 2000, he was a Senior Research Fellow with the Department of Computing, Hong Kong Polytechnic University, Kowloon. He is currently a Professor of computer science with the Harbin Institute of Technology Shenzhen Graduate School. His research interest includes artificial intelligence, machine learning, computational linguistics, and Chinese information processing.

Qiang Wang received the B.S. degree in computer science and technology and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2002 and 2007, respectively.

He is with the New Search Research and Development Department, Baidu Company, Beijing, China. His research interests include the theories and methods for question answering, machine learning, and text mining.

Mr. Wang participated in The First International Joint Conference on Natural Language Processing (IJCNLP2004) and gave an oral presentation.