Learning a Concept-Based Document Similarity Measure

Lan Huang, David Milne, Eibe Frank, and Ian H. Witten
Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.
E-mail: {lh92, dnk2, eibe, ihw}@cs.waikato.ac.nz

Document similarity measures are crucial components of many text-analysis tasks, including information retrieval, document classification, and document clustering. Conventional measures are brittle: They estimate the surface overlap between documents based on the words they mention and ignore deeper semantic connections. We propose a new measure that assesses similarity at both the lexical and semantic levels, and learns from human judgments how to combine them by using machine-learning techniques. Experiments show that the new measure produces values for documents that are more consistent with people's judgments than people are with each other. We also use it to classify and cluster large document sets covering different genres and topics, and find that it improves both classification and clustering performance.

Introduction

Accurate assessment of the topical similarity between documents is fundamental to many automatic text-analysis applications, including information retrieval, document classification, and document clustering. Choosing a good similarity measure is no less important than is choosing a good document representation (Hartigan, 1975). Commonly used techniques such as the cosine and Jaccard metrics rely on surface overlap: To be related, documents must use the same vocabulary.

These existing measures treat words as though they are independent of one another, which is unrealistic. In fact, words are not isolated units but always relate to each other to form meaningful structures and to develop ideas. When reading, our thoughts constantly utilize relations between words to facilitate understanding (Altmann & Steedman, 1988). Without resolving word-level redundancies (i.e., synonymy) and ambiguities (i.e., polysemy), a similarity computation cannot accurately reflect the implicit semantic connections between words.

The alternative we investigate in this article is to use concepts instead of words to capture the topics of documents, by creating a concept-based document representation model. "Concepts" are units of knowledge (International Organization for Standardization, 2009), each with a unique meaning. They have three advantages over words as thematic descriptors. First, they are less redundant because synonyms such as *U.S.* and *United States* unify to the same concept. Second, they disambiguate terms such as *apple* and *jaguar* that have multiple meanings. Third, semantic relations between concepts can be defined, quantified, and taken into account when computing the similarity between documents—for example, a document discussing *endangered species* may relate to one on *environmental pollution* even though they may have no words in common.

The value of concepts and their relations has been recognized and exploited in many text-processing tasks, including information retrieval (Milne, Witten, & Nichols, 2007), semantic analysis (Mihalcea, Corley, & Strapparava, 2006), document classification (Gabrilovich & Markovitch, 2005), and document clustering (Hu et al., 2008; Huang, Milne, Frank, & Witten, 2008). Some authors even have enriched document similarity measures based on lexical or conceptual overlap with semantic relations between concepts (Hu et al., 2008; Mihalcea et al., 2006); however, this was done in an ad hoc fashion, and the best way to employ such rich semantic knowledge remains unknown. To establish a more principled approach, we use supervised machine-learning techniques to determine how to combine concepts and their semantic relations into a document similarity measure that reflects human judgment of thematic similarity.

We evaluate the learned measure in two types of tasks. First, we compare it with human judgments of document similarity in terms of the consistency it achieves with human raters. Empirical results show that it produces values for documents that are more consistent with people's judgments than people are with each other. Second, we use it to classify and cluster documents from different sources. Empirical results show that it outperforms existing overlap-based similarity measures by obtaining better classification accuracy and document clusters with greater cohesion. Our results

Received October 20, 2011; revised February 24, 2012; accepted February 24, 2012

© 2012 ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22689

provide strong support for the learned measure's generality: It can be used effectively on documents with different topics and genres, from different subject domains, and with varying lengths.

The next section reviews related work. Then, we introduce our framework and its key components: how concepts in documents can be identified and used to represent them, and how their semantic relations can be quantified and exploited to calculate document similarity. Next, we present a new approach for automatically learning the document similarity measure from human judgments. Finally, we describe the evaluations performed and discuss the results.

Related Work

The standard document representation technique is the vector space model (Salton, Wong, & Yang, 1975). Each document is expressed as a weighted high-dimensional vector, the dimensions corresponding to individual features such as words or concepts. When words are used, the result is called the *bag-of-words* model. It is brittle because of redundancy, ambiguity, and orthogonality; the first because synonyms are not unified, the second because no account is taken of polysemy—one word can have different meanings—and the third because semantic connections between words are neglected, which not only encompasses the synonymy and polysemy relations but extends to the more general sense of two words being semantically related.

Alternative features such as phrases (Caropreso, Matwin, & Sebastiani, 2001), term clusters (Slonim & Tishby, 2000), and statistical topic models (Blei, Ng, & Jordan, 2003; Hofmann, 1999) have been proposed to solve these problems. However, phrases, being sequences of words, also can be ambiguous, although they are usually more specific than are single-word terms. For example, *access point* usually refers to a device used to connect to a wireless network, yet it also can mean a rocky point on the Anvers Island of Antarctica.

Term-clustering and topic-modeling techniques seek groups or combinations of terms that are strongly associated with one another in a given document collection, each cluster or combination presumably representing a latent topic hidden in the documents. Their effectiveness depends heavily on the input data. In addition, it is hard to interpret which topic a term cluster or combination represents (Hofmann, 1999), let alone connect the topics. Therefore, these techniques cannot easily be generalized to fresh data, particularly documents with previously unseen terms and ones from different document collections.

Concepts—units of knowledge—provide a neat solution to these problems. Each one represents a unique meaning and is thus unambiguous; because of this, semantic relations between concepts can be defined and quantified to address the orthogonality problem. Concepts make more succinct descriptors than do words.

Concepts, organized and structured according to the relations among them, form a concept system. Given the standard definition of concepts as units of knowledge, encyclopedias such as *Britannica* (Britannica, 2011) and Wikipedia (Wikipedia, 2011) are promising sources of concept knowledge. They provide extensive coverage of almost every branch of knowledge, with a particular focus on factual explanations of the concepts (Hartmann & James, 1998). *Britannica* is available only commercially, so we focus on the freely accessible Wikipedia. Some resources such as the Medical Subject Headings (MeSH) and Agrovoc are domain-dependent. Research on and applications of these systems are usually restricted to processing texts from that domain (Bloehdorn & Hotho, 2004; Zhu, Zeng, & Mamitsuka, 2009). Thus, they are not considered in this article, although the techniques developed here can be directly applied to such resources.

Wikipedia is a collaboratively developed, online encyclopedia in which each article succinctly describes a single topic that we treat as a "concept." The English version contains 3.7-million articles.[1] Because of its open accessibility and comprehensive world knowledge, Wikipedia has been extensively and effectively exploited to facilitate better understanding of documents. Studies have shown that Wikipedia-based concept representations are more effective than are word vectors when assessing the semantic relatedness between documents (Gabrilovich & Markovitch, 2007; Yeh, Ramage, Manning, Agirre, & Soroa, 2009), and have been applied successfully to information retrieval (Milne et al., 2007; Potthast, Stein, & Anderka, 2008), text classification (Gabrilovich & Markovitch, 2005; Wang & Domeniconi, 2008), and document clustering (Hu et al., 2008; Huang et al., 2008).

Lexical resources also have been exploited to identify concepts in running text. These provide information about individual words rather than general conceptual knowledge (Gabrilovich & Markovitch, 2009). In particular, WordNet (Miller, 1995) is a lexical ontology of common English word knowledge expressed in terms of concepts called *synonym sets* (*synsets*), maintained by experts at Princeton University. The most recent version (3.0) contains about 118,000 concepts. It also encodes semantic relations among concepts, such as *generic* (hypernymy) and *partitive* relations (meronymy). Concept representations based on WordNet have been utilized to quantify semantic relatedness between documents (Mihalcea et al., 2006; Mohler & Mihalcea, 2009), and in information retrieval (Gonzalo, Verdejo, Chugur, & Cigarran, 1998; Voorhees, 1998), text classification (Gabrilovich & Markovitch, 2004; Scott & Matwin, 1999), and document clustering (Hotho, Staab, & Stumme, 2003; Recupero, 2007).

Concept-based document representations solve the redundancy and the ambiguity problems, but are still basically orthogonal. To address this problem, some have expanded the representation to incorporate concepts that are absent from a document, but closely related to ones that it mentions (Bloehdorn & Hotho, 2004; Gabrilovich &

[1] Statistics are based on the snapshot of September 8th, 2011.

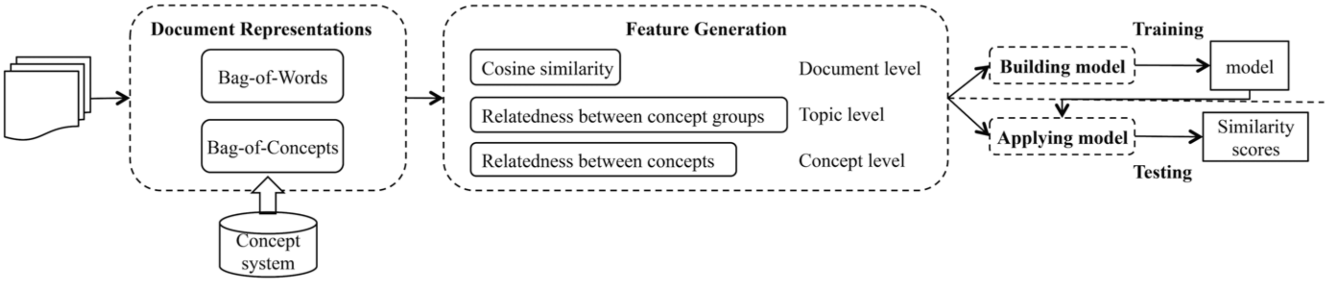

FIG. 1. The process of creating and applying our document similarity measure.

Markovitch, 2005; Recupero, 2007; Yeh et al., 2009), and others only have considered relations that are pertinent to the documents currently being compared (Hu et al., 2008). The decisions governing which relations should be considered and how are usually ad hoc. For example, Bloehdorn and Hotho (2004) expanded to concepts that are more general than those mentioned in the document, and restrained the expansion to be within a certain depth in a hierarchy. Hu et al.'s (2008) system considers several relations, including hierarchical and associative relations, each restricted to a certain range, and the formula for combining them is determined empirically through experimental trials.

In contrast, our work takes explicit account of semantic relations between concepts, in a principled way. Related methods in the literature include explicit semantic analysis (ESA; Gabrilovich & Markovitch, 2005) and its successor graph-based ESA (ESA-G; Yeh et al., 2009), both of which index documents with Wikipedia concepts based on full-text analysis. ESA indexes a document with Wikipedia articles that have certain surface overlap with it. ESA-G enriches ESA with hyperlink structure information by using an iterative random walk (Page, Brin, Motwani, & Winograd, 2009) over Wikipedia's hyperlink graph that is initialized with the Wikipedia concepts assigned to a document by ESA. Because they require processing the full text of Wikipedia articles, they are computationally more expensive than is our method, which does not involve full-text analysis. Later in the article, we compare our measure with these techniques.

We use both WordNet and Wikipedia to identify concepts in documents and to relate different concepts. Both are domain-independent, yet different techniques are required because they have distinct structures and characteristics. We will explain how each is used to identify concepts in free-text documents after we introduce our framework in the following section.

Framework

Figure 1 illustrates the general process of creating and applying our document similarity measure. Given a document collection, we first list all the possible document pairs. Given each pair, the first step creates two independent representations by extracting words and concepts from the documents. The feature-generation step takes the representations as input, extracts features that describe the resemblance between the two documents at different levels, and outputs a feature vector. The feature vectors for different document pairs are used to build the similarity measure in the training phase, and the resulting model is then applied to previously unseen document pairs to predict their thematic similarity.

The next section explains the document representations. Several features involve measuring the semantic relatedness between concepts; thus, we will first describe the measures that we use for WordNet and Wikipedia, and then introduce the features.

Document Representation

Documents are represented at the lexical and semantic levels by the *words* and *concepts* they contain. This creates two independent representations, called *bag-of-words* and *bag-of-concepts*, respectively. To create the former, documents are segmented into tokens based on white space, paragraph separators, and punctuation marks. Then all words are extracted and stemmed (Porter, 1980), stop words are removed, and the number of occurrences of each word is counted.

To create the bag-of-concepts representation, the concepts in the document are identified. First, an index vocabulary is extracted from each concept system (Wikipedia and WordNet) whose entries associate concepts with lists of expressions that could be used to refer to it in running text (Huang et al., 2008). For Wikipedia, the expressions come from the redirects and anchor phrases that point to a Wikipedia article. For WordNet, they are the synonyms in a synset. For example, WordNet associates the concept "a machine for performing calculations automatically" with 6 expressions (*computer*, *computing machine*, *computing device*, *data processor*, *electronic computer*, and *information processing system*) whereas Wikipedia associates it with more than 100, from synonyms such as *computer systems* to common spelling errors such as *computar*.

Concepts mentioned in a document are identified in two steps: *candidate identification* and *sense disambiguation*. In the first step, all word sequences up to the maximum length of the index vocabulary are extracted, provided that they do not cross boundaries such as paragraph separators. Each sequence is matched against the vocabulary. A positive match connects it to the concept or concepts associated with that expression, which constitute the set of concept candidates. For example, *pluto* generates (at least) three concept candidates: the dwarf planet in the solar system, the cartoon character, and the Roman god of the underworld. The second step disambiguates the intended meaning of a polysemous term and retains only the concept that represents this meaning.

For Wikipedia, the disambiguation process establishes how closely a concept candidate relates to its surrounding context and chooses the most highly related as the intended sense (Milne & Witten, 2008a). For WordNet, we simply choose the most common sense of a concept as its intended sense because experimentation has shown this to be the most effective method (Hotho et al., 2003; Huang, 2011). In either case, the outcome is the bag-of-concepts representation of the document, comprising the set of concepts it mentions, synonyms having been mapped to the same concept and polysemous terms having been disambiguated as described earlier, along with a count of the number of occurrences of each concept.

Semantic Connections Between Documents

Researchers have long been aware of the redundancy, ambiguity, and orthogonality problems (Hartigan, 1975); however, they cannot be solved using the bag-of-words model. The concept-based model rectifies the situation. The previous section explained how concepts address redundancy and ambiguity; now, we focus on orthogonality. More specifically, we quantify how closely concepts relate to each other, and integrate this into a document similarity measure. As a result, documents do not have to mention the same words or concepts to be judged similar.

Concept Relatedness Measure

Measuring semantic relatedness between concepts is a challenging research problem in its own right and has been studied extensively using both WordNet and Wikipedia (Gabrilovich & Markovitch, 2007; Leacock & Chodorow, 1997; Milne & Witten, 2008b; Resnik, 1995; Strube & Ponzetto, 2006). There are three general requirements for a concept relatedness measure to be applicable in our framework. First, it should be accurate, an appropriate measure of accuracy being consistency with human judgments of relatedness. Second, it should apply to all members of the concept system simply because any concept could be encountered in practice. Third, it should be symmetric. Although asymmetry may be desirable in some tasks (Tătar, Şerban, Mihiş, & Mihalcea, 2009), the tasks to which we apply the relatedness measure predominantly use symmetric relationships. Relatedness values also need to be normalized to the range from 0 (*completely unrelated*) to 1 (*synonymous*).

For WordNet concepts, we use Leacock and Chodorow's (1997) path-length measure LCH. For Wikipedia, we use Milne and Witten's (2008b) hyperlink-structure measure: the Wikipedia link-based measure (WLM). They both satisfy all three requirements. They are either more accurate than are the alternatives in terms of consistency with human judgment or are as accurate, but significantly more efficient (Milne & Witten, 2008b; Strube & Ponzetto, 2006).

LCH utilizes WordNet's concept taxonomies, and defines semantic distance as the number of nodes along the shortest path between two concepts, normalized by the depth of the taxonomy. Formally, the relatedness between concepts A and B is defined as

$$LCH(A, B) = -\log \frac{length(A, B)}{2D}$$

where $length(A, B)$ is the number of nodes along the shortest path between A and B, and D is the maximum depth of the taxonomy. If A and B belong to different taxonomies (e.g., A is a noun, and B is a verb), or either concept does not exist in any taxonomy (e.g., A is an adjective), the relatedness is set to zero.

WLM has two components: modeling incoming and outgoing hyperlinks, respectively. Given two Wikipedia articles A and B, denote the sets of hyperlinks found within them by A_{out} and B_{out}, and the sets of hyperlinks that are made to them by A_{in} and B_{in}. WLM's first component uses the cosine measure between A_{out} and B_{out}:

$$WLM_{out}(A, B) = \frac{\sum_{l \in A_{out} \cup B_{out}} w(l, A) \times w(l, B)}{\sqrt{\sum_{l \in A_{out}} w(l, A)^2} \times \sqrt{\sum_{l \in B_{out}} w(l, B)^2}}.$$

Here, $w(l, A)$ is the weight of a link l with respect to article A, which is 0 if $l \notin A$ and $\log \frac{|W|}{|T|}$ otherwise, where $|W|$ is the total number of articles in Wikipedia and $|T|$ the number that link to the target of l. This resembles inverse document frequency weighting (Manning, Raghavan, & Schütze, 2008). Incoming links are modeled after the *normalized Google distance* (Cilibrasi & Vitányi, 2007). Formally,

$$WLM_{in}(A, B) = 1 - \frac{\max(\log|A_{in}|, \log|B_{in}|) - \log|A_{in} \cap B_{in}|}{\log(|W|) - \log(\min(|A_{in}|, |B_{in}|))}$$

where $A_{in} \cap B_{in}$ denotes the set of hyperlinks that link to both A and B. WLM computes overall relatedness as the average of these two components.

Context Centrality

Next, we integrate concept relatedness into a full measure of document similarity. A key notion is the *centrality* of a concept with respect to a given context, where a "context" is the set of concepts in a document. Centrality indicates the concept's relevance to the context, and we use it to enrich the overlap-based measure.

DOI: 10.1002/asi

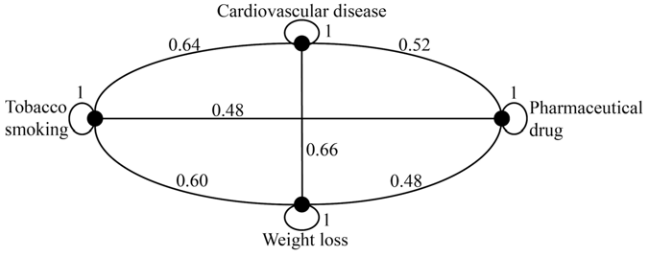

FIG. 2. Example concept graph.

Concepts and their connections are represented by a weighted undirected graph whose vertices are concepts and whose edges connect pairs of concepts, weighted by their relatedness. Figure 2 shows an example. When two concepts have zero relatedness, we create an edge with zero weight. We also create an edge from each vertex to itself, with a weight of 1, to cope with the situation when the graph has only one concept. The concepts themselves are weighted, too, the weights being either binary—0 and 1 indicating the concept's absence and presence, respectively—or a numeric score reflecting how often the concept is mentioned in the context. We call these the *binary* and *weighted* schemes.

We compute the context centrality of each concept within the document by calculating the average edge weight of its vertex in the graph. Formally, denote the weight of concept c in a set of concepts C by $w(c, C)$ (either binary or numeric, as noted earlier). The centrality of c with respect to the context C is defined as

$$CC(c, C) = \frac{\sum_{c_j \in C} rel(c, c_j) \times w(c_j, C)}{\sum_{c_j \in C} w(c_j, C)},$$

where $rel(c, c_j)$ is the relatedness between c and c_j. Context centrality is normalized between 0 and 1, and higher values indicate that the concept is closer to the *center* of the graph (i.e., more closely related to the context).

Learning Document Similarity

We use a total of 17 features to characterize document similarity, representing four different aspects: *overall similarity*, *context centrality*, *strongest connection*, and *concept groups*. The first bases similarity on the entire bag-of-concepts; the second and third utilize the strength of semantic connections beyond the documents' surface forms; and the fourth takes into account the relations between "topics," which we define as groups of closely related concepts.

The learned similarity measure takes as input a pair of documents and produces a score between 0 and 1, the former indicating that the concepts in the documents are completely different and the latter that they are identical. The measure works in three steps: *document representation*, *feature extraction*, and *similarity calculation*. The first step generates the bag-of-words and bag-of-concepts representations discussed earlier, the second calculates feature values that characterize the thematic resemblance between documents, and the third uses a model to compute similarity according to the feature values. This model is first built from training data—document pairs with their thematic similarity rated by human raters—and then applied to previously unseen document pairs. It encodes how the features should be combined to best model human judgment.

Next, we introduce the features. We then describe how the model is built from training data and is applied to fresh documents.

Features

Each of the four aspects mentioned earlier consists of several features, reflecting the various perspectives the aspect encompasses. Each feature is expressed with one or two *attributes* that correspond to dimensions in the vector representing the document pair. The 17 features result in the 25 attributes listed in Table 1. These attributes comprise the vector that describes the similarity between a pair of documents.

Overall similarity. The first feature type computes the similarity between documents based on the overall similarity of the entire bag of words or concepts that they mention. For bags of words, we use the cosine measure, and for bags of concepts, we develop a new, enriched measure that takes into account semantic relations between the concepts. This generates two features, *CosineWords* and *EnrichedConcepts* (F1 and F2 in Table 1), each of which corresponds to one attribute.

TABLE 1. Features used for learning document similarity.

Level	Aspect	ID	Feature	No. of attributes
Document	Overall similarity	F1	CosineWords	1
		F2	EnrichedConcepts	1
Concept	Context centrality	F3	MaxLocal	2
		F4	MinLocal	2
		F5	AvgLocal	2
		F6	SDLocal	2
		F7	MaxRelative	2
		F8	MinRelative	2
		F9	AvgRelative	1
		F10	SDRelative	2
	Strongest connection	F11	MaxRel	1
		F12	MaxNORel	1
Topic	Concept groups	F13	AvgGroupSize	2
		F14	MaxGroupRel	1
		F15	MinGroupRel	1
		F16	AvgGroupRel	1
		F17	SDGroupRel	1
Total		F1–F17		25

D1	By giving up **smoking**, **losing weight**, and becoming more active people can reduce their risk of **cardiovascular disease** two to three-fold, which largely outweighs the risks of taking the **medications**.	**smoking** → *Tobacco smoking* **losing weight** → *Weight loss* **cardiovascular disease** → *Cardiovascular disease* **medications** → *Pharmaceutical drug*
D2	In **the UK**, there are 2 million people affected by **angina**: the most common symptom of **coronary heart disease**. **Angina pectoris**, commonly known as **angina**, is severe **chest pain** due to **ischemia** (a lack of **blood**, hence a lack of **oxygen** supply) of the **heart muscle**.	**the UK** → *United Kingdom* **angina** → *Angina pectoris* **coronary heart disease** → *Coronary heart disease* **chest pain** → *Chest pain* **ischemia** → *Ischemia* **blood** → *Blood* **oxygen** → *Oxygen* **heart muscle** → *Cardiac muscle*

FIG. 3. Documents about *smoking* and *health* respectively.

Cosine similarity measure. The cosine measure calculates the similarity between two documents as the cosine of the angle between their corresponding word vectors (Salton et al., 1975). Formally, if $\overrightarrow{d_A}$ and $\overrightarrow{d_B}$ are the word vectors of documents d_A and d_B, their similarity is computed as

$$\text{cosine}(d_A, d_B) = \frac{\overrightarrow{d_A}\cdot\overrightarrow{d_B}}{\left|\overrightarrow{d_A}\right|\cdot\left|\overrightarrow{d_B}\right|} = \frac{\sum_{t\in V} w(t, d_A)\times w(t, d_B)}{\sqrt{\sum_{t\in V} w(t, d_A)^2}\sqrt{\sum_{t\in V} w(t, d_B)^2}},$$

where $w(t,d_A)$ is the weight of word t in d_A—that is, the *tf* × *idf* weight, based on the number of occurrences of t in d_A—and V denotes the size of the vocabulary of the document collection. Evidence from various applications has shown that this formula effectively measures interdocument similarity (Lee, Pincombe, & Welsh, 2005; Rorvig, 1999; Willett, 1983).

Like many other measures, the cosine measure does not take connections between features into account. Thus, we only apply it to the word-based representation. Despite its limitations, this feature (F1) contributes to the learned measure's robustness, especially in the extreme (and extremely rare) case where no concepts are detected in the input document.

Semantically enriched similarity. One way of enriching document similarity with semantic relations is to expand each document's representation to include new concepts based on those it already mentions: both more generic concepts such as hypernyms of existing ones (Bloehdorn & Hotho, 2004; Recupero, 2007) and closely related concepts (Hu et al., 2008). For example, Figure 3 shows two documents, the Wikipedia concepts identified in them (italicized, on the right), and the phrases in the documents that evoke the concepts (in bold). The documents have no concepts in common, yet the first mentions *cardiovascular disease* and the second mentions *coronary heart disease*, both of which belong to the same Wikipedia category, *cardiovascular diseases*. The two documents could be related by expanding both representations to include this common category. Of course, the expansion must be restricted somehow—perhaps to concepts within a certain range.

For example, the concept *smoking*, which is literally mentioned in Document D1, might be expanded to include hypernyms such *as addiction and habits*, and closely related concepts such as *tobacco*, *cigarette*, and *nicotine*. However, most of these expanded concepts are irrelevant for connecting Document D1 with Document D2, which discusses coronary heart disease.

An alternative approach to enriching document similarity is to focus on the comparison itself, and take account of the context that the comparison provides. The orthogonality problem when comparing two documents can be addressed by enriching each document with the concepts that have been identified in the other: here, enriching D2 with the four concepts in D1, and enriching D1 with the eight concepts in D2. We utilize the measures of concept relatedness explained previously to determine the weights of the enriched concepts.

Given two documents, we enrich each by adding all the new concepts that are identified in the other. The weight of each newly added concept is based on both its most closely related concept in the document to which it has been added and its centrality with respect to that document. Formally, given documents d_A and d_B with concept sets C_A and C_B, we first enrich C_A with concepts from d_B that are not mentioned in d_A. For each such concept c_e ($c_e \in C_B$ and $c_e \notin C_A$), the first component—its strongest connection with C_A—is denoted by c_e^A; that is, $c_e^A = \max_{c\in C_A} rel(c_e, c)$, and the second component is its centrality with C_A: $CC(c_e, d_A)$. The enriched concept c_e's weight in d_A is

$$w_e(c_e, d_A) = w(c_e^A, d_A)\times rel(c_e^A, c_e)\times CC(c_e, d_A),$$

where $w(c_e^A, d_A)$ is c_e's most related concept c_e^A's weight in d_A, which also is weighted with the *tf* × *idf* scheme based on its occurrence frequencies, and $rel(c_e^A, c_e)$ is their relatedness. Document d_B is enriched in the same way with concepts from d_A that are not mentioned in d_B. Then, the cosine measure is used with the enriched representations.

Both components of a newly added concept's weight—its strongest semantic connection and its context centrality with the document—are plausibly necessary. The former represents the most likely strength of the connection that the concept makes between the two documents while the latter adjusts it according to the concept's importance in the document to which it has been added.

Context centrality. The second group of features characterizes the distribution of the context centrality values of concepts in each document. We calculate centrality with respect to two distinct contexts: (a) the one surrounding a concept, which reflects how central it is to the document that mentions it, and (b) the context provided by the comparison document, which reflects the concept's relevance to the comparison itself. We call these *local* and *relative* centrality, respectively.

Four statistics are used to describe the overall distribution of centrality values, *minimum*, *maximum*, *average*, and

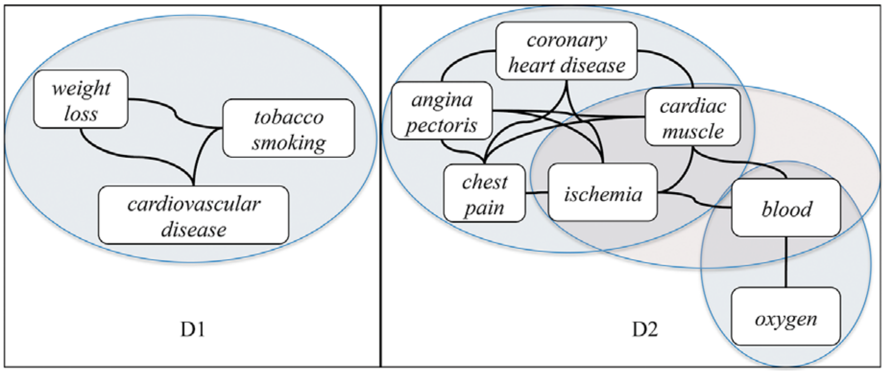

FIG. 4. Concept groups in the documents in Figure 3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

standard deviation, and these are applied to both local and relative centrality to yield features F3 to F6 and F7 to F10 in Table 1. The distribution of local centrality reveals the cohesiveness of a document while the distribution of relative centrality characterizes the overall semantic relatedness between the documents. For example, if two documents share similar topics, a considerable proportion of their concepts should have high relative centrality, resulting in a large average and a small standard deviation.

The first two features—minimum and maximum—are trivial to obtain. The average centrality is the mean of the centrality values of all concepts, and the standard deviation is computed as

$$s = \sqrt{\frac{1}{|C|-1}\sum_{c_i \in C}(CC(c_i, C) - \overline{CC})^2}$$

where $|C|$ is the number of concepts in the context, $\underline{CC}(c_i, C)$ is concept c_i's centrality with respect to C, and $\overline{CC}$ is the average context centrality of all concepts in C. Each feature yields two attributes, except for the average relative centrality (F9), which is symmetric.

Strongest connection. The centrality features assess relations between one concept and a set of concepts. For example, maximum centrality identifies which concept has the strongest *overall* relatedness to *all* concepts in a group. The third group in Table 1 concerns one-to-one relations, which also provide useful information about document similarity. There are two such features: the maximum relatedness between single concepts in the documents (*MaxRel*, F11), and the maximum relatedness between concepts that appear in one document but not the other (*MaxNORel*, F12). The two are the same unless the documents have at least one concept in common, in which case *MaxRel* = 1.

For example, the strongest connection between D1 and D2 in Figure 3 is between *cardiovascular disease* and *coronary heart disease*, whose relatedness value is 0.71. Because the two documents have no concepts in common, *MaxRel* = *MaxNORel* = 0.71 for this example.

Concept groups. Concepts mentioned in the same document are not only related but also can form their own structures: Closely related concepts are often used together when describing a topic with which they are all associated. For example, a document explaining *oil spill* might mention alternative references to oil (e.g., *petroleum*, *gasoline*, *diesel*), some oil companies (e.g., *Shell* and *BP*), and oil's influence on species such as *seabirds* and *marine mammals*. These three groups (*oil spill*, etc.; oil companies; and wildlife species) each represent a more detailed aspect of the document's topic. Documents that share similarity in any aspect are somewhat similar to the document in question, and those that mention all three aspects are even more alike.

To capture this effect, we cluster concepts according to their relatedness to each other, combining closely related ones into the same group and separating those with tenuous links into different groups. Each group, like the three discussed earlier, reflects a topic or a subtopic mentioned in the document. From these, we generate features that describe interdocument relations at the *topic* level, which is intermediate between the document and concept levels examined previously.

Specifically, concepts are clustered to form *cliques*—complete subgraphs— to make the topics (or subtopics) as coherent as possible. Again, documents are modeled by weighted undirected graphs with concepts as vertices. Unlike the graphs used to model context centrality, which connect every concept to all others, here only those whose relatedness exceeds a certain threshold (e.g., 0.5 is used throughout this article) are connected. The maximal cliques of this graph give the concept groups we seek. Every pair of concepts assigned to the same group exceeds this threshold, and no other concept can be added to any of these groups.

For example, Figure 4 shows the groups with at least two concepts identified from D1 and D2 (Figure 3) with a relatedness threshold of 0.5. D1 contains just one group, and it is closely related to only the first of the three groups in D2. *Blood* appears in two of D2's groups because *ischemia*, *cardiac muscle*, and *oxygen* are insufficiently related for the groups to be merged.

DOI: 10.1002/asi

Concepts that cannot be assigned to any group can either form a singleton—a group by itself—or be ignored. We call these the *full* and *strict* schemes, respectively. For example, D1 has one group in the *strict* scheme and two groups in the *full* scheme. Full schemes capture every aspect of a document, even the unimportant ones. This favors situations when two documents are highly similar: Even their less important aspects can be alike and strongly related. In contrast, strict schemes highlight the most prominent aspects of a document and can reduce computation overhead by avoiding relatedness computation between singleton groups.

Several features can be derived from these concept groups. One is the average group size for each document (F13 in Table 1). As with centrality, we use the maximum, minimum, average, and standard deviation statistics to characterize the distribution of relatedness between groups (F14–F17).

The relatedness between two concept groups is the weighted average of the relatedness between their member concepts. Formally, let $G_A = (\sigma_1, . . ., \sigma_q)$ and $G_B = (\sigma_1, . . ., \sigma_q)$ be the concept groups (σ) identified from documents d_A and d_B. The relatedness between σ_h from d_A and σ_l from d_B is calculated as

$$rel(\varsigma_h, \varsigma_l) = \frac{\sum_{c_i \in \varsigma_h} \sum_{c_j \in \varsigma_l} w(c_i, d_A) \times w(c_j, d_B) \times rel(c_i, c_j)}{|\varsigma_h| \times |\varsigma_l|},$$

where $|\sigma|$ refers to the size of group σ and is calculated as $\sum_{c \in \varsigma} w(c, d)$. Here, $w(c, d)$ is concept c's weight in document d that produces σ, and is either 1 or 0 to indicate the concept's presence or absence (the *binary* version), or a score based on its number of occurrences (the *weighted* version). The average group relatedness is the mean of every possible pair of concept groups weighted by each group's size:

$$grouprel(d_A, d_B) = \frac{\sum_{\varsigma_h \in G_A} \sum_{\varsigma_l \in G_B} rel(\varsigma_h, \varsigma_l) \times |\varsigma_h| \times |\varsigma_l|}{\sum_{\varsigma_h \in G_A} |\varsigma_h| \sum_{\varsigma_l \in G_B} |\varsigma_l|}.$$

If no concept group is found for a document (in the *strict* scheme), the average group relatedness is set to −1 to differentiate this from the case where none of the groups are related, in which case $grouprel(d_A, d_B)$ is zero.

Training Data

Our strategy is to build a model that uses these features to predict document similarity; for this, we need training data. Unfortunately, little data on manually rated thematic document similarity are available, and we know of only one data set with a substantial number of human raters, referred to as *HE50* (Lee et al., 2005; Pincombe, 2004).

HE50 consists of 50 short news documents from August 2002, selected from a group of articles taken from the Australian Broadcasting Corporation's news mail service. The documents are quite short—between 51 and 126 words—and contain a total of 1,583 distinct words after case-folding. Assessments of word distribution show that they are normal English documents (Pincombe, 2004). The documents were paired in all possible ways, generating 1,225 pairs (excluding self pairs).

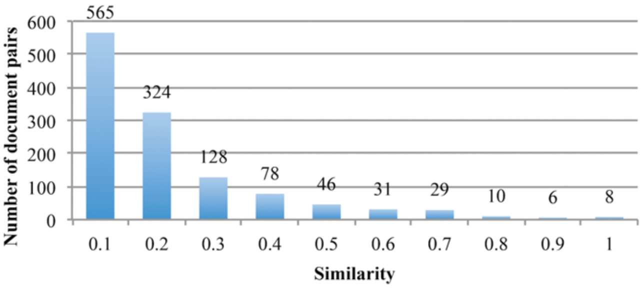

FIG. 5. Distribution of averaged human ratings in the HE50 data set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The judges were 83 students from the University of Adelaide. Document pairs were presented in random order, and the order of documents within each pair also was randomized. Students rated the pairs on an integer scale from 1 (*highly unrelated*) to 5 (*highly related*), each pair receiving 8 to 12 human judgments. Judgments were averaged and normalized to [0,1]. Figure 5 shows the distribution of the normalized ratings.

Inter-labeler consistency is assessed in terms of Pearson's linear correlation coefficient. Lee et al. (2005) showed that the raters' judgments are quite consistent throughout the task and with each other: On average, raters have 0.6 correlation with one another. This data set has become the benchmark for evaluating document similarity measures. One aim is to make automated measures as consistent with humans as humans are among themselves, and this interrater consistency serves as a baseline.

Regression Algorithms

We use regression algorithms to build a model that makes numeric predictions based on numeric values. We have experimented with several such algorithms: linear regression, support vector machines for regression (SVMreg; Smola & Schölkopf, 2004), the Gaussian process for regression (Rasmussen & Williams, 2006), and these four algorithms applied with forward stagewise additive modeling for regression (Hastie, Tibshirani, & Friedman, 2009). Performance is measured in terms of consistency with human judgments. The best results were achieved with support vector machines using the radial basis function kernel; incorporating additive regression improves performance only slightly, and the improvement is not statistically significant. Thus, all the results that follow were obtained with the SVMreg regressor alone (with $\varepsilon = 1.0E{-}12$, $C = 1.0$, and $\gamma = 0.01$ for the RBF kernel). All attributes are first

TABLE 2. Performance on the HE50 data set.

Method	Pearson's correlation coefficient	
Baselines	Interrater (Lee et al., 2005)	0.6
	Bag-of-words with cosine measure (Lee et al., 2005)	0.42
	Latent Semantic Indexing (Lee et al., 2005)	0.6
	Explicit Semantic Analysis (ESA; Gabrilovich & Markovitch, 2005)	0.72
	Graph-based Explicit Semantic Analysis (ESA-G; Yeh et al., 2009)	0.77
Our method		0.808

standardized to have zero mean and unit variance, except the class attribute—the average rated similarity.

Evaluation Strategy

We evaluate the learned measure against human judgment, and we also evaluate it in specific applications. The former investigates whether the measure is able to predict thematic document similarity as consistently as do humans, and also explores the effectiveness and predictive ability of the two concept systems—WordNet and Wikipedia—and of the individual features. The latter is an important addition to evaluation against human judgment (Budanitsky & Hirst, 2001). It tests the measure's effectiveness in different scenarios by applying it to different data sets and to document classification and clustering, which are both tasks that require a document similarity measure. This is important because the training data set (HE50) is tiny, and any measure learned from it will overfit unless it generalizes to other tasks and documents.

Evaluation Against Human Judgments

Like other researchers, we use Pearson's linear correlation coefficient to measure the consistency between the predicted similarity and the gold standard—the average similarity as judged by human raters. The coefficient for two samples X and Y with n values and means $\bar{X}$ and $\bar{Y}$ is defined as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Table 2 summarizes results in the literature on this data set. Lee et al. (2005) found the cosine measure with the bag-of-words representation to yield a correlation of 0.42 with the gold standard, with only trivial differences between different similarity measures (including Jaccard). Their best result was achieved using latent semantic analysis on a larger collection of 364 documents, also from Australian Broadcasting Corporation news. Document vectors are transformed to a new feature space consisting of the latent topics identified in the larger set, and the cosine measure is used with the new vectors. As Table 2 shows, this technique is as consistent with an average human rater as the raters are with themselves. None of the bag-of-words similarity measures approach this level. Furthermore, research has shown that estimates of interrater consistency based on partial document sets can be overoptimistic (Westerman, Cribbin, & Collins, 2010), which bolsters our method's performance.

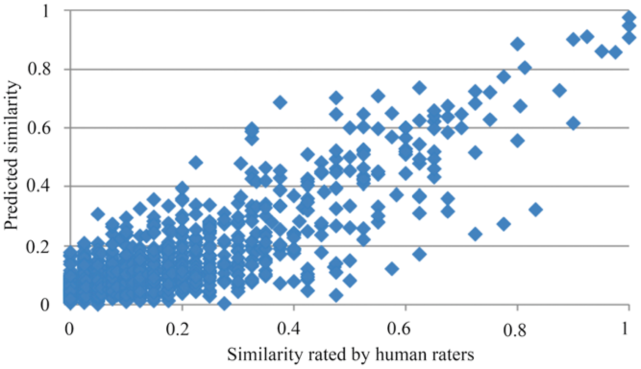

FIG. 6. Correlation between predicted similarity and that of human raters. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Gabrilovich and Markovitch (2005) and Yeh et al. (2009) also reported results on the HE50 data set. Both their systems, ESA and ESA-G, represent documents as vectors of Wikipedia concepts and use the cosine measure; both yield a greater correlation with human raters than does the average interrater agreement. These are the state-of-the-art results on this data set. These two methods and the interrater consistency comprise the three baselines for assessing our learned similarity measure.

Evaluation Setup

Each document is represented in three different ways: a bag-of-words and bags-of-concepts based on Wikipedia and WordNet. The average bag sizes per document are 37.1 words, 13.1 Wikipedia concepts, and 39.2 WordNet concepts, for a total of 1,187 words, 492 Wikipedia concepts, and 1,201 WordNet concepts.

All results reported next are averaged over five independent runs of stratified 10-fold cross-validation to help reduce the possibility of overfitting the learned measure to the tiny HE50 data set. In each run, the regression algorithm was trained on 90% of the document pairs (1,102 examples) and tested on the remaining 10% (123 examples). To indicate the predictive capability of the learned model on new data, performance was measured on the held-out test set. Paired corrected resampled t tests (Nadeau & Bengio, 2003) were used to establish statistical significance at a confidence level of 0.05.

Overall Consistency With Human Judgment

Our best model achieved an average correlation of 0.808 with the human ratings, ranging from 0.66 to 0.88. Figure 6 plots the similarity predicted by the best learned measure,

DOI: 10.1002/asi

TABLE 3. Predictive value of individual features (using Wikipedia concepts).

			Pearson's correlation			
Aspect	ID	Feature	Binary		Weighted	
Overall similarity	F1	CosineWords	0.57			
	F2	EnrichedConcepts	0.717		0.710	
Context centrality	F3	MaxLocal	–0.039		–0.001	
	F4	MinLocal	–0.043		0.038	
	F5	AvgLocal	0.374		0.045	
	F6	SDLocal	0.022		0.004	
		LocalCentralityCombined	*0.155*		*0.174*	
	F7	MaxRelative	0.691		0.685	
	F8	MinRelative	0.703		0.707	
	F9	AvgRelative	0.327		0.320	
	F10	SDRelative	0.679		0.657	
		RelativeCentralityCombined	*0.725*		*0.711*	
		CentralityCombined	*0.774*		*0.759*	
Strongest connection	F11	MaxRel	0.62			
	F12	MaxNORel	0.643			
		MaxRelatednessCombined	*0.688*			
			Strict	Full	Strict	Full
Concept groups	F13	AvgGroupSize	0.176	0.137	0.176	0.137
	F14	MaxGroupRel	0.655	0.481	0.655	0.489
	F15	MinGroupRel	0.002	0.001	0.001	0.001
	F16	AvgGroupRel	0.664	0.608	0.674	0.665
	F17	SDGroupRel	0.474	0.618	0.451	0.624
		GroupRelatednessCombined	*0.7*	*0.689*	*0.703*	*0.718*
Overall	F1–F17		*0.808*	*0.799*	*0.805*	*0.801*

using the Wikipedia-concept-based document representation, against that of human raters, and the ideal case would be a diagonal line from (0,0) to (1,1). The high correlation is apparent: Its value of 0.808 exceeds both the interrater consistency and the state-of-the-art result obtained by Yeh et al. (2009). The upper right and lower left corners show that the learned measure agrees particularly well with human judgment on highly similar and highly dissimilar document pairs. Most points are concentrated at the lower left corner because in this small data set, most documents have different topics. In fact, each document has an average of only 3.1 other documents whose manually rated similarity exceeds 0.5.

WordNet Versus Wikipedia

The learned measure's correlation with human judgment is only 0.611 using the WordNet-based concept model, suggesting that Wikipedia concepts and the WLM concept relatedness measure are more effective than are their WordNet counterparts in this task, and that Wikipedia's world knowledge is more relevant to thematic document similarity than is WordNet's lexical knowledge. While the WordNet model did little better than people (0.6, see Table 2), it roundly outperforms the bag-of-words representation using cosine similarity (0.42).

Predictive Power of Individual Features

The predictive power of an individual feature is assessed from the performance of the regression model learned from that feature alone. Table 3 shows the individual predictability of the 17 features when Wikipedia concepts are used, along with some combinations.[2]

The difference between F1 and F2 indicates that the new representation is more discriminative than the bag-of-words model with the usual cosine measure—which is remarkable because there are more than twice as many distinct words as there are concepts. Furthermore, F2's improvement over F1 is statistically significant.

The *context centrality* section of Table 3 shows that relative centrality is far more informative than is local centrality, with a dramatic difference in the performance of both individual features and their combinations (*LocalCentralityCombined* and *RelativeCentralityCombined*).

Local centrality focuses on the quality of an individual document. It is a measure of how homogenous a document is, of how much it follows a single thread. Relative centrality, in contrast, focuses on the relation between two documents, on whether they talk about closely related topics. Intuitively, relative centrality was always going to be the more useful measure. After all, our end goal is to measure the relatedness between two documents. This is borne out in Table 3, where minimum relative centrality on its own approaches the performance of ESA (0.7 vs. 0.72 correlation).

However, local centrality can still make a contribution. Imagine comparing three news stories, two of which discuss

[2] For results on WordNet concepts, see http://cs.waikato.ac.nz/~lh92/learned.html.

the cargo ship Rena running aground off the coast of New Zealand, and the third story places this in the broader context of other threats to local wildlife. All three documents are related, but for the third one, this is diluted by the presence of other threads of discussion. This distinction is captured by local centrality. Consequently, the *Centrality-Combined* measure (with correlations of 0.774 and 0.759) performs better than does relative centrality alone (with correlations of 0.725 and 0.711).

The "strongest connection" section of Table 3 shows that both these features strongly predict document similarity regardless of whether the binary or weighted representation is used. We do not consider the weakest concept connection—the minimum concept relatedness—because it barely correlates with human judgment (in fact, it is usually zero).

The distinction between the strict and full schemes—whether stray concepts that cannot be assigned to any groups are treated as singletons—affects all features in the "concept groups" section of Table 3, so their results are shown separately. All results were obtained using a relatedness threshold of 0.5 for creating concept cliques. The first feature—the averaged size of concept groups in each document (F13)—does not involve a concept's number of occurrences in a document, so the binary and weighted schemes produce the same result. The minimum relatedness between concept groups (F15) contributes little because even documents with similar topics usually mention some unrelated concepts, giving it a value close to zero. The average size of concept groups is not effective either, especially when compared with the other three features (F14–F17). As with the local centrality features, this is probably because it describes characteristics of the document itself while the others target relations between documents.

All features except the first (F1, the bag-of-words representation) involve concepts and utilize the relatedness between them. If no concept is identified in a document, all their values are missing, and the model relies on F1 to make a prediction.

Removing Less Informative Features

Five features stand out as significantly less informative than the others: the three local centrality features (F3, F4, and F6), the average concept group size (F13), and the minimum relatedness between concept groups (F15). Excluding these reduces the space from 17 features and 25 attributes to 12 features and 17 attributes. Table 4 compares the performance of the model trained before and after removing these features.

Discarding uninformative features is advantageous in most cases, although the differences are not statistically significant. Table 4 also shows that stray concepts are better ignored rather than treated as singleton clusters: The strict schemes outperform the full schemes, and the improvements in both cases are statistically significant. Yet, the difference between the two strict schemes—binary and weighted—is not significant.

TABLE 4. Performance of the reduced feature set on HE50.

	Binary		Weighted	
Features (No. of attributes)	Strict	Full	Strict	Full
Full: F1–F17 (25)	0.808	0.799	0.805	0.801
Reduced: F1–F2, F5, F7–F12, F14, F16–F17 (17)	0.808	0.8	0.807	0.8

TABLE 5. Statistics of the four experimental data sets.

Data set	Categories	Documents	Category size
SmallReuters	30	1,658	55.3
NewsSim3	3	2,938	979.3
NewsDiff3	3	2,780	926.7
Med100	23	2,256	98.1

Evaluation in Document Classification and Clustering

In addition to the previous evaluation, we tested the learned measure in two applications: document classification and document clustering. Both benefit from an accurate measure of interdocument similarity.

In this evaluation, the full HE50 data set is used to train the regression model instead of using 10-fold cross-validation as before. This is safe because we are now testing the outcome on previously unseen data. The model is built with the *binary strict* scheme and the reduced feature set—12 features and 17 attributes.

Test Data

We create four data sets from standard corpora whose thematic components are already labeled for evaluating classification and clustering performance.[3] The first three, SmallReuters, NewsSim3, and NewsDiff3, contain short news articles and newsgroup posts covering diverse topics while the last, Med100, contains medical papers from MEDLINE and is thus domain-specific. Each data set has different properties, topic domains, and difficulty levels. Table 5 shows summary statistics.[4]

Evaluation of Document Classification

Document classification is the task that automatically classifies a document into categories that are already known. There exist many classification methods, and we test the

[3] The original Reuters, 20Newsgroups, and OHSUMed collection are available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, http://people.csail.mit.edu/jrennie/20Newsgroups/, and http://ir.ohsu.edu/ohsumed/.

[4] More details are available at http://cs.waikato.ac.nz/~lh92/learned.html.

DOI: 10.1002/asi

TABLE 6. Performance of the learned measure in document classification.

	Word cosine	Learned measure	
SmallReuters	0.881	0.924*	4.9%
NewsSim3	0.860	0.833*	−3.2%
NewsDiff3	0.971	0.976*	0.5%
Med100	0.515	0.591*	14.8%

*Statistically significant improvement/degradation.

learned measure with instance-based classifiers (Aha, Kibler, & Albert, 1991), which predict the class of a test instance based on its closest (i.e., most similar) neighbor(s) in the training set, and thus require an accurate interdocument similarity measure. We use the standard *k*-nearest-neighbor classifier, denoted by *k*NN.

Classification performance is measured with the F_1 measure widely used in information retrieval. Let $\Omega = (\omega_i, \ldots, \omega_w)$ denote the set of classes in the data set. Given a class ω_i, *precision*(ω_i) and *recall*(ω_i) are defined as

$$P(\omega_i) = \frac{\text{\# documents in class } \omega_i \text{ that are classified as } \omega_i}{\text{\# documents that are classified as } \omega_i},$$

$$R(\omega_i) = \frac{\text{\# documents in class } \omega_i \text{ that are classified as } \omega_i}{\text{\# documents in class } \omega_i}.$$

The F_1 measure is the harmonic mean of the two:

$$F_1 = \frac{2P(\omega_i)R(\omega_i)}{P(\omega_i) + R(\omega_i)}.$$

The overall F_1 measure is the weighted sum of F_1 over all classes $\{L_1, \ldots, L_m\}$ in the data set:

$$F_1 = \sum_{\omega_i \in \Omega} \frac{|\omega_i|}{N} F_1(\omega_i),$$

where $|\omega_i|$ is the number of documents in class ω_i, and N is the total number of documents in the data set.

Overall performance. We performed 10 runs of 10-fold cross-validation with *k*NN on each data set, and report the average classification performance. The best number of nearest neighbors in the range 1 to 10 that yields the greatest F1 score was determined using leave-one-out cross-validation (Aha et al., 1991).

Table 6 compares the learned measure (using Wikipedia concepts) with the baseline method—the usual bag-of-words representation—and shows, using a paired *t* test, whether the difference in performance is statistically significant. It achieves significant improvement on two data sets—SmallReuters and Med100. One possible reason why it fails to show improvement on the two newsgroup data sets is that NewsSim3 and NewsDiff3 contain only three classes and have many training examples—over 800 per class in each fold. This makes it more likely for a test instance to share considerable surface overlap with one of the training examples.

Varied training set. To investigate the impact of the likelihood that testing documents share surface overlap with training documents on classification performance, we varied the proportion of training examples from 5 to 95% in increments of 5%, and used the remaining examples for testing. Each of the 19 trials was run 10 times, with different training sets. The order of the training examples was randomized, and the best number of neighbors was sought as described earlier.

Figure 7 shows the clear advantage of the learned measure for small training sets, particularly with NewSim3 (Figure 7c) and NewsDiff3 (Figure 7d). This suggests that it might be helpful when there is little overlap (of words or concepts) between the training and test examples because the semantic connections between concepts can effectively relate documents with similar topics, but different surface forms.

Advantages of the learned measure are more consistent on the SmallReuters and the Med100 data sets. This is because the categories in the NewsSim3 and NewsDiff3 data sets are much larger: Each category has about 900 documents on average. This means that taking 5% of the documents for training will result in about 45 training documents for each category, which is equivalent in size to taking 80% of the SmallReuters data set and 45% of the Med100 data set as training data. This indicates that the learned measure is particularly beneficial for problems with small training sets, which is important in practice because obtaining labeled training data is often expensive.

Evaluation of Document Clustering

Clustering is another important technique in practical data mining. Document clustering is the task that automatically analyzes the relations among documents and organizes them to form thematically coherent structures—clusters of documents that share similar topics. We tested two commonly used algorithms: hierarchical agglomerative clustering with group-average-link (Manning et al., 2008) and *k*-means (Hartigan, 1975). Performance is measured in terms of goodness of fit with the existing categories in the data set using the normalized mutual information (NMI) measure. For each data set, the number of clusters being sought equals the number of categories. Each cluster is labeled by the most frequent category in that cluster.

Let $\Phi = \{\rho_1, \ldots, \rho_k\}$ and $\Omega = \{\omega_1, \ldots, \omega_m\}$ denote the set of clusters and categories, respectively. The NMI measure is defined as

$$NMI(\Phi, \Omega) = \frac{I(\Phi; \Omega)}{[H(\Phi) + H(\Omega)]/2},$$

where I is the mutual information between the set of clusters and the set of categories, formally:

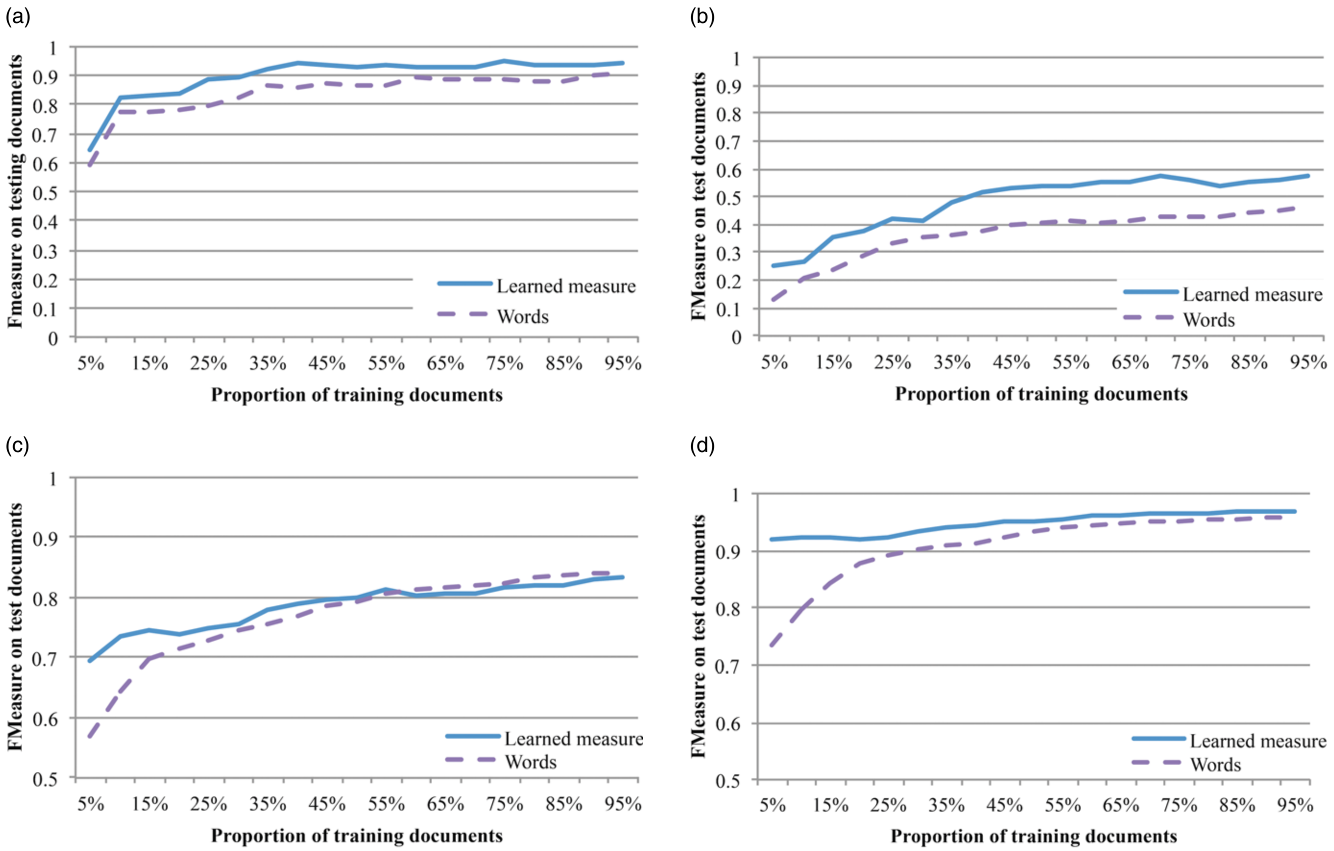

FIG. 7. Learning curves for the four data sets: SmallReuters (a), Med100 (b), NewsSim3 (c), and NewsDiff3 (d). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$I(\Phi;\Omega)=\sum_{\rho_i}\sum_{\omega_j}\frac{|\rho_i\cap\omega_j|}{N}\log\frac{N|\rho_i\cap\omega_j|}{|\rho_i||\omega_j|}.$$

Here, N is the total number of documents in the data set, H is the entropy, and $H(\Phi)$ is defined as

$$H(\Phi)=-\sum_{\rho_i}\frac{|\rho_i|}{N}\log\frac{|\rho_i|}{N},$$

and the same for $H(\Omega)$. Unlike other common measures for cluster quality, such as purity and inverse purity, NMI is independent of the number of clusters and can be used to measure the overall structural fitness of a clustering with respect to the categories (Manning et al., 2008).

Note that the standard *k*-means algorithm represents a cluster by its centroid, and this representation differs from a normal document—for example, it has a nonzero value for every word or concept mentioned in any document in that cluster. Instead, we represented a cluster by its members, and measure its similarity to another document or group of documents by taking the average similarity with all member documents.

Table 7 shows the results. The learned similarity measure is very effective, and outperforms the baseline on every data set. This is particularly remarkable in three respects. First, the training data set is tiny—it only contains 50 documents—yet the learned measure can be effectively applied to larger corpora. Second, documents in the training data set are significantly shorter than are those in the experimental data sets—37 words compared to over 100, on average—yet the learned measure remains effective. Third, documents in the training data set come from different sources and cover different topics from those in the clustering data set, which demonstrates that the learned measure is both generic and robust.

TABLE 7. Performance of the learned measure in document clustering.

	Hierarchical		*k*-means	
	Word cosine	Learned measure	Word cosine	Learned measure
SmallReuters	0.588	0.696*	0.687	0.792*
Med100	0.276	0.365*	0.209	0.348*
NewsSim3	0.027	0.167*	0.008	0.298*
NewsDiff3	0.180	0.613*	0.149	0.724*

*Statistically significant improvement.

The learned measure gains most on the NewsDiff3 data set. This is because Wikipedia concepts are thematically dense descriptors—they provide topic-related information—

DOI: 10.1002/asi

while some words merely reflect lexical features and are common to documents with different topics. For example, adverbs and adjectives such as "significantly" and "beautiful" rarely provide topic-related information. When the documents have very different topics, concepts can retain the main thematic features of a document and discard the unimportant lexical features, thus making the distinction even more prominent, which facilitates clustering.

Computational Complexity

Computational complexity mainly comes from two steps: the creation of document representations (i.e., bag-of-words and bag-of-concepts) and application of the learned measure based on the representations. The first step is in general linear to the lengths of the two documents. Let w and c denote the average number of words and concepts found in a document. The overall complexity of the second step is quadratic to the number of concepts and linear to the number of words. The cosine similarity measure is linear: $O(w)$ and $O(c)$. Both the local and relative centrality measures are quadratic: $O(c^2)$. For the concept group features, we use the Bron–Kerbosch algorithm (Bron & Kerbosch, 1973) to find the maximal cliques, which has an $O(c)$ time complexity for sparse graphs. Furthermore, concept relatedness can be cached to speed up the computation of centrality measures. In practice, it usually takes a couple of seconds to predict for a pair from the experimental data sets.

Computational complexity of the training phase also contains two parts: the creation of document representations and training the regression model. The latter depends on the computational complexity of the regression method that is used, but is generally negligible with the amount of labeled training data that we deal with. As an indicative result, in our experiments training took less than 1 second with 17 attributes and 1,225 training documents.

Conclusion

We have developed a novel method for learning an inter-document similarity measure from human judgment. It overcomes the redundancy, ambiguity, and orthogonality problems that plague traditional methods of computing document similarity by using concepts instead of words as document descriptors and taking the semantic connections between concepts into account. The measure predicts similarity more consistently with average human raters than human raters do between themselves, and also outperforms the current state of the art on a standard data set. Furthermore, both the features used for describing document similarity and the machine-learning technique used to build the model are generic. The resulting measure applies to documents from different sources and topic domains, and improves performance when classifying and clustering these documents: in the former case in particular when only a small amount of training data is present per class.

It is no surprise that concepts are better thematic descriptors of text than are words. During the 1980s, researchers began to develop formal concept systems such as WordNet to facilitate computer processing of natural language text, but success was limited, and the bag-of-words model still prevails in practice. With the advent of Web 2.0 and the birth of collaboratively constructed and informal, yet comprehensive, online encyclopedias such as Wikipedia, the use of concepts and their relations began to attract increasing attention as a replacement for words and other lexical features.

Our results provide strong support for why people should be encouraged to abandon the old models and methods. We have developed an alternative that is based on concepts, and have demonstrated that it is general and effective. The new method is not confined to the classification and clustering tasks tested here but applies wherever text must be analyzed and organized according to its topics.

The results of this research are available in the form of an open source toolkit called *knowledge assisted text organization algorithms* (Katoa) that implements the concept-based document representations generator using WordNet and Wikipedia, and the similarity measure learned from human judgment.[5]

References

Aha, D.W., Kibler, D., & Albert, M.K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37–66.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. Cognition, 30(3), 191–238.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Bloehdorn, S., & Hotho, A. (2004). Boosting for text classification with semantic features. In Proceedings of the Sixth International Workshop on Knowledge Discovery on the Web (WebKDD) (pp. 149–166). Berlin: Springer-Verlag.

Britannica. (2011). The Encyclopaedia Britannica (2010 copyright). Available at: http://www.britannica.com/

Bron, C., & Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph. Communications of the ACM, 16(9), 575–577.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the Second Workshop on WordNet and Other Lexical Resources of the North American Chapter of the Association for Computational Linguistics. Available at: ftp://ftp.db.toronto.edu/pub/gh/Budanitsky+Hirst-2001.pdf

Caropreso, M.F., Matwin, S., & Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A.G. Chin (Ed.), Text databases and document management: Theory and practice (pp. 78–102). Hershey, PA: IGI Global.

Cilibrasi, R.L., & Vitányi, P.M. (2007). The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 19(3), 370–383.

Gabrilovich, E., & Markovitch, S. (2004). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the 21st International Conference on Machine Learning (pp. 41–48). New York: ACM Press.

[5] Katoa is available from http://katoa.sf.net

Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (pp. 1048–1053). San Francisco: Kaufmann.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In Proceedings of the 20th International Joint Conference on Artifical Intelligence (pp. 1606–1611). San Francisco: Kaufmann.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 34(1), 443–498.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL'98) Workshop on Usage of WordNet for NLP (pp. 38–44). Available at: http://acl.ldc.upenn.edu/W/W98/W98-0705.pdf

Hartigan, J.A. (1975). Clustering algorithms. New York: Wiley.

Hartmann, R.R., & James, G. (1998). Dictionary of lexicography. London: Routledge.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2009). Boosting and additive trees. New York: Springer.

Hofmann, T. (1999). Probablistic latent semantic indexing. In Proceedings of the 22nd International Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval (pp. 50–57). New York: ACM Press.

Hotho, A., Staab, S., & Stumme, G. (2003). WordNet improves text document clustering. In Proceedings of the Semantic Web Workshop at the 26th annual International Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) Conference on Research and Development in Information Retrieval. New York: ACM Press (pp. 541–544).

Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., et al. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In Proceedings of the 31st annual International Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) Conference on Research and Development in Information Retrieval (pp. 179–186). New York: ACM Press.

Huang, A. (2011). Concept-based text clustering. Unpublished doctoral dissertation, University of Waikato, Hamilton, Waikato, New Zealand.

Huang, A., Milne, D., Frank, E., & Witten, I.H. (2008). Clustering documents with active learning using Wikipedia. In Proceedings of the Eighth IEEE International Conference on Data Mining (pp. 839–844). Washington, DC: IEEE Computer Sciety.

International Organization for Standardization. (2009). ISO-704: Terminology work—Principles and methods (3rd ed.). Geneva, Switzerland: Author.

Leacock, C., & Chodorow, M. (1997). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), WordNet: An electronic lexical database (pp. 265–283). Cambridge, MA: MIT Press.

Lee, M.D., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In Proceedings of the 27th annual Conference of the Cognitive Science Society (pp. 1254–1259). Mahwah, NJ: Erlbaum.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the 21st National Conference on Artificial Intelligence (pp. 775–780). Menlo Park, CA: AAAI Press.

Miller, G.A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39–41.

Milne, D., & Witten, I.H. (2008a). Learning to link with Wikipedia. In Proceedings of the 17th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (pp. 509–518). New York: ACM Press.

Milne, D., & Witten, I.H. (2008b). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceedings of the Advancement of Artificial Intelligence Workshop on Wikipedia and Artificial Intelligence (AAAI 2008) (pp. 25–30). Menlo Park, CA: AAAI Press.

Milne, D., Witten, I.H., & Nichols, D.M. (2007). A knowledge-based search engine powered by Wikipedia. In Proceedings of the 16th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (pp. 445–454). New York: ACM Press.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 567–575). Stroudsburg, PA: Association for Computational Linguistics.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab. Available at http://ilpubs.stanford.edu:8090/422/.

Pincombe, B. (2004). Comparison of human and latent semantic analysis (LSA) judgements of pairwise document similarities for a news corpus. Canberra: Australia Government Department of Defense, Information Sciences Laboratory, Intelligence, Surveillance and Reconnaissance Division.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In Proceedings of the 30th European Conference on Advances in Information Retrieval Research (pp. 522–530). Berlin: Springer-Verlag.

Rasmussen, C.E., & Williams, C.K.I. (2006). Gaussian processes for machine learning. Cambridge, MA: MIT Press.

Recupero, D.R. (2007). A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. Information Retrieval, 10(6), 563–579.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 448–453). San Francisco: Kaufmann.

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science and Technology, 50(8), 639–651.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automaic indexing. Communications of the ACM, 18(11), 613–620.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In Proceedings of the 16th International Conference on Machine Learning (pp. 379–388). San Francisco: Kaufmann.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In Proceedings of the 23rd Annual International Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) Conference on Research and Development in Information Retrieval (pp. 208–215). New York: ACM Press.

Smola, A.J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222.

Strube, M., & Ponzetto, S.P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence (pp. 1419–1424). Menlo Park, CA: AAAI Press.

Tătar, D., & Şerban, G., Mihiş, A., & Mihalcea, R. (2009). Textual entailment as a directional relation. Journal of Research and Practice in Information Technology, 41(1), 53–64.

Voorhees, E.M. (1998). Using WordNet for text retrieval. In C. Fellbaum (Ed.), WordNet: An electronic lexical database (pp. 285–304). Cambridge, MA: MIT Press.

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using Wikipedia. In Proceedings of the 14th annual International Association for Computing Machinery Special Interest Group on Knowledge Discovery in Data (ACM SIGKDD) Conference on Knowledge Discovery and Data Mining (pp. 713–721). New York: ACM Press.

Westerman, S.J., Cribbin, T., & Collins, J. (2010). Human assessments of document similarity. Journal of the American Society for Information Science and Technology, 61(8), 1535–1542.

Wikipedia. (2011). Wikipedia. http://www.wikipedia.org/ (Last access March 26, 2011).

Willett, P. (1983). Similarity coefficients and weighting functions for automatic document classification: An empirical comparison. International Classification, 10(3), 138–142.

Yeh, E., Ramage, D., Manning, C.D., Agirre, E., & Soroa, A. (2009). WikiWalk: Random walks on Wikipedia for semantic relatedness. In Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing (pp. 41–49). Stroudsburg, PA: Association for Computational Linguistics.

Zhu, S., Zeng, J., & Mamitsuka, H. (2009). Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics, 25(15), 1944–1951.