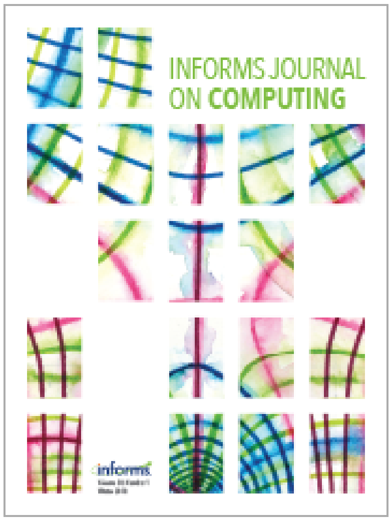
This article was downloaded by: [216.73.216.140] On: 08 June 2026, At: 13:16
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Learning Context-Sensitive Domain Ontologies from Folksonomies: A Cognitively Motivated Method

Raymond Y. K. Lau, J. Leon Zhao, Wenping Zhang, Yi Cai, Eric W. T. Ngai

To cite this article:
Raymond Y. K. Lau, J. Leon Zhao, Wenping Zhang, Yi Cai, Eric W. T. Ngai (2015) Learning Context-Sensitive Domain Ontologies from Folksonomies: A Cognitively Motivated Method. INFORMS Journal on Computing 27(3):561-578. https://doi.org/10.1287/ijoc.2015.0644

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2015, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

INFORMS Journal on Computing
Vol. 27, No. 3, Summer 2015, pp. 561–578
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.2015.0644
© 2015 INFORMS

Learning Context-Sensitive Domain Ontologies from Folksonomies: A Cognitively Motivated Method

Raymond Y. K. Lau, J. Leon Zhao, Wenping Zhang
Department of Information Systems, College of Business, City University of Hong Kong, Kowloon, Hong Kong SAR
{raylau@cityu.edu.hk, jlzhao@cityu.edu.hk, wzhang23-c@my.cityu.edu.hk}

Yi Cai
School of Software Engineering, South China University of Technology, Guangzhou, Guangdong, China, ycai@scut.edu.cn

Eric W. T. Ngai
Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong SAR,
mswtngai@polyu.edu.hk

Ontology is the backbone of the Semantic Web, helping users search for relevant resources from the Web of linked data. The existing context-free mapping approach between tags and concepts fails to address the problems of social synonymy and social polysemy when ontologies are induced from folksonomies. The novel contributions of this paper are threefold. First, grounded in the cognitively motivated category utility measure, a novel basic-level concept mining algorithm is developed to construct semantically rich concept vectors to alleviate the problem of social synonymy. Second, contextual aspects of ontology learning are exploited via probabilistic topic modeling to address the problem of social polysemy. Third, a novel context-sensitive domain ontology learning algorithm that combines link- and content-based semantic analysis is developed to identify both taxonomic and associative relations among concepts. To the best of our knowledge, this is the first successful research that exploits a cognitively motivated method to learn context-sensitive domain ontologies from folksonomies. By using the Open Directory Project ontology as a benchmark, we examined the effectiveness of the proposed algorithms based on social annotations crawled from three different folksonomy sites. Our experimental results show that the proposed ontology learning system significantly outperforms the best baseline system by 13.83% in terms of taxonomic F-measure. The practical implication of our research is that high-quality ontologies are constructed with minimal human intervention to facilitate concept-driven retrieval of linked data and the knowledge-based interoperability among enterprises.

Keywords: folksonomies; ontology learning; machine learning; artificial intelligence; knowledge management
History: Accepted by Alexander Tuzhilin, (former) Area Editor for Knowledge and Data Management; received February 2013; revised March 2014; accepted January 2015. Published online September 23, 2015.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

1. Introduction

Social annotation services, also called folksonomies, allow a large number of users simultaneously and collaboratively to exchange metadata about Web resources by using freely chosen tags (i.e., key words). Well-known examples of social annotation services include delicious.com for sharing Web pages, Instagram.com for exchanging photos, and citeulike.org for sharing academic publications. The notion of folksonomies originally refers to folk-generated taxonomies. Compared to traditional hand-crafted taxonomies that are built with a controlled vocabulary and maintained by a limited number of experts, folk-generated taxonomies are the by-products of leveraging the "collective intelligence" of a Web scale to efficiently induce the emerging "semantics" of Web resources. Learning ontologies from folksonomies is much cheaper than the traditional hand-crafted method because there is no need to hire domain experts to build concept hierarchies of different domains.

Recently, researchers have paid more attention to the approach of learning ontologies from folksonomies because they believe that folksonomies are more reliable knowledge sources than free texts (Liu et al. 2010, Strohmaier et al. 2012, Tang et al. 2009, Trabelsi et al. 2010). Ontology is a formal specification of concepts and their relationships (Gruber 1993). Ontology either captures a taxonomy of concepts (i.e., a lightweight ontology) or comprises a taxonomy and the axioms characterizing it (i.e., a heavyweight ontology) (Wong et al. 2012). More specifically, domain ontologies capture the precise knowledge of specific application domains (Lau et al. 2009, Zouaq and Nkambou 2009). Ontologies are very useful to enhance various applications such as Semantic Web, electronic commerce, knowledge management, and enterprise interoperability (Missikoff and Taglino 2004, Panetto et al.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

2012). The recent trend of the Semantic Web is toward the construction of linked data (Bizer et al. 2009). However, searching for relevant information from linked data of a Web scale is a nightmare. Ontologies can help users efficiently search for relevant information from linked data via concept-based information retrieval (Yan et al. 2011). This paper focuses on a cognitively motivated methodology of learning lightweight domain ontologies from folksonomies.

The problems of synonymy (i.e., different words referring to the same concept) and polysemy (i.e., the same word carrying different meanings) arise in many natural language processing (NLP) applications. However, these problems escalate in a social tagging environment because there are a huge number of users simultaneously referring to the same concept using different tags (i.e., the social synonymy problem), and a relatively short tag (usually one or two words) is applied to describe a variety of concepts by many annotators (i.e., the social polysemy problem) (Mika 2007, Schmitz 2006). For instance, though many users employ the tag “cloud” to annotate “service-oriented” resources, an equal number of users applies the same tag to annotate “photography”-related resources in folksonomies such as delicious.com. Existing folksonomy-based ontology learning methods are weak in addressing the social synonymy and social polysemy problems since they often adopt a context-free, one-to-one mapping approach between tags and concepts (Heymann and Garcia-Molina 2006, Liu et al. 2010, Trabelsi et al. 2010). Moreover, existing methods lack the cognitive foundation to induce concepts from social annotations even though social tagging is a kind of human conceptualization process.

Since social annotations reflect how humans naturally conceptualize objects (e.g., Web resources), an effective folksonomy-based ontology learning method should be grounded in human cognition, that is, imitating how people learn and apply concepts in a folksonomy environment. In the field of cognitive psychology, empirical results have shown that people tend to categorize objects using concepts that carry the appropriate level of semantics (i.e., basic-level concepts) (Rosch et al. 1976, Tanaka and Taylor 1991). For example, when people see a British Shorthair named Kitty, they tend to say that this is a “cat” rather than a “mammal.” Basic-level concepts refer to the prominent concepts of a domain that are most naturally recognized by humans (Tanaka and Taylor 1991). To identify basic-level concepts of a domain, psychologists have developed a cognitively motivated measure named category utility (Gluck and Corter 1985). Empirical tests have revealed that these basic-level concepts often carry relatively high category utilities (Gluck and Corter 1985). Accordingly, it is intuitively attractive to apply category utility to extract the basic-level concepts that social taggers use to categorize Web resources. From this standpoint, our proposed cognitively motivated ontology learning method is quite different from most existing methods.

This paper presents our research on mining implicit semantics from folksonomies to build lightweight domain ontologies. Our proposed methodology, named context-sensitive domain ontology learning (CSDOL), exploits theories and techniques from the fields of cognitive psychology, machine learning, NLP, and information retrieval to improve the effectiveness of ontology learning. The novel contributions of our research work are as follows: (1) grounded in the cognitively motivated category utility measure, we develop a novel basic-level concept mining algorithm, which simulates the human conceptualization process, for extracting semantically rich concept vectors from social annotations to alleviate the social synonymy problem; (2) “contexts” of ontology learning are exploited by applying a probabilistic topic model to dynamically extract contextual aspects from domain-specific corpora to address the social polysemy problem; (3) we develop a novel context-sensitive domain ontology learning algorithm that combines link- and content-based semantic analysis to construct ontologies supporting multiple inheritance. To the best of our knowledge, this is the first successful research of designing a cognitively motivated method to learn context-sensitive domain ontologies from folksonomies. The practical implication of our research is that high-quality ontologies are constructed with minimal human intervention to facilitate the retrieval of linked data on the Web and knowledge-based interoperability among enterprises.

2. Related Work

2.1. Learning Ontologies from Folksonomies

Schmitz (2006) applied a subsumption-based approach to extract tag subsumption network for the construction of faceted ontologies. Zhou et al. (2007) developed a hierarchical divisive clustering algorithm to learn concept hierarchies from social annotations. Ponzetto and Strube (2007) employed NLP techniques to build a taxonomy of concepts from Wikipedia. Mika (2007) proposed a tripartite graph model to represent the emerging actor-concept-instance semantics captured by folksonomies. Tang et al. (2009) examined a probabilistic generative model for the construction of concept taxonomy. The main idea is that a tag is mapped to multiple topics (concepts). Our proposed ontology learning method also considers actor-sensitive semantics arising in a collaborative social tagging environment. However, our method differs from the aforementioned approaches in that domain-specific contexts are exploited to disambiguate concepts and relations.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

Daud et al. (2010) proposed the actor-concept-instance-topic model to capture emerging semantics in folksonomies. More specifically, latent dirichlet allocation (LDA) was applied to extract the association relations among actors, concepts, and instances captured in a tripartite graph. Our proposed method differs in that LDA is applied to extract the contextual aspects of a domain to disambiguate both concepts and relationships. Liu et al. (2010) employed formal concept analysis and the measures of rule support and rule confidence to construct a tag subsumption graph based on social annotations. Trabelsi et al. (2010) applied triadic concept analysis and utilized the Stanford syntactic dependence parser to learn nontaxonomic relationships among tags retrieved from folksonomy sites. Our proposed methodology differs from these methods in that it supports a semantically rich approach of many-to-many mapping between tags and domain concepts.

A semiautomatic method was proposed to refine a domain ontology using social annotations extracted from an e-learning environment (Gasevic et al. 2011). By applying four folksonomy induction algorithms to five different social tagging data sets, Strohmaier et al. (2012) found that folksonomy induction algorithms that leveraged tag network analysis (e.g., degree centrality) to estimate tag generality were more effective than the traditional clustering-based algorithms. After extracting frequently occurring adjective-noun pairs (i.e., concepts) from folksonomies, Borth et al. (2013) utilized an emotion lexicon to construct a sentiment ontology; the ontology was then applied to predict the polarities of visual contents. Our proposed method leverages a more sophisticated ontology learning method to extract both taxonomic and associative relations among concepts. A comparison between our proposed ontology learning method and existing methods is depicted in Table 1. It is shown that most existing methods employ a direct one-to-one mapping approach between tags and concepts, and the concept learning techniques are not context-sensitive in most cases. Above all, none of the existing method is designed based on rigorous cognitive foundation.

2.2. Learning Ontologies from Free Texts

The seminal work of Hearst (1992) applied predefined lexicosyntactic patterns to learn taxonomies of concepts from free texts. Sanderson and Croft (1999) proposed a document-based subsumption induction method to extract taxonomic relations among terms. We extend the term-based subsumption method to design a concept-based approach that utilizes both link- and content-based semantic analysis. Roussinov and Zhao (2003) leveraged Web documents retrieved from a search engine as the context to learn similarity relations among terms. Cimiano et al. (2005) presented a taxonomy learning algorithm that extracted concept hierarchies from text corpora according to formal concept analysis. The fuzzy ontology generation framework (FOGA) extended the formal concept analysis approach by applying the notion of fuzzy sets (Tho et al. 2006). Lau et al. (2009) applied latent semantic analysis to prune the concept space and developed a fuzzy subsumption function to learn domain ontologies from messages posted to an e-learning environment. For the proposed CSDOL methodology, latent semantic analysis is applied to extract contextual aspects from a domain-specific corpus to disambiguate concepts. Zouaq and Nkambou (2009) proposed a lexicosyntactic method, named TEXCOMON, to automatically learn both taxonomic and nontaxonomic relations from free texts. Wei et al. (2010) applied probabilistic topic models to extract concept hierarchies from free texts. Tao et al. (2011) developed a semiautomatic personalized ontology construction method by using manually constructed knowledge bases such as the Library of Congress subject headings and personal bookmarks. The proposed CSDOL method differs from the aforementioned

Table 1 Comparison Among Folksonomy-Based Ontology Learning Methods

Method	Community sensitivity	Domain sensitivity	Tag-concept mapping	Taxonomic rel. learning	Association rel. learning	Learning technique	Cognitive foundation
Schmitz (2006)	Yes	No	One-to-one	Yes	No	Term subsumption	No
Zhou et al. (2007)	No	No	One-to-one	Yes	No	Divisive clustering	No
Ponzetto and Strube (2007)	No	No	One-to-one	Yes	No	Dependency parsing	No
Mika (2007)	Yes	No	One-to-one	Yes	Yes	Community linking	No
Tang et al. (2009)	No	No	Many-to-many	No	Yes	Topic modeling	No
Daud et al. (2010)	Yes	No	One-to-one	No	Yes	Topic modeling	No
Liu et al. (2010)	No	No	One-to-one	Yes	No	Formal concept analysis	No
Trabelsi et al. (2010)	No	No	One-to-one	No	Yes	Dependency parsing	No
Gasevic et al. (2011)	No	Yes	One-to-one	No	Yes	Mutual information	No
Strohmaier et al. (2012)	No	No	One-to-one	Yes	No	Tag network analysis	No
Borth et al. (2013)	No	No	One-to-one	No	Yes	Term co-occurrence	No
Our method—CSDOL	Yes	Yes	Many-to-many	Yes	Yes	Category utility, tag generality context-sensitive subsumption	Yes

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

approaches in that it supports context-sensitive mining of domain ontologies from folksonomies instead of free texts.

2.3. Evaluation of Ontologies

It has been stressed that evaluation of ontology learning algorithms is often a challenging research problem by itself (Strohmaier et al. 2012, Wong et al. 2012). In the context of evaluating the performance of four folksonomy induction algorithms, Strohmaier et al. (2012) proposed reference-based, human-based, and application-oriented assessments for the ontologies generated by folksonomy induction algorithms. Zavitsanos et al. (2011) proposed a probabilistic concept matching method to alleviate the weakness of the simple string matching-based ontology evaluation method. Jupiter et al. (2010) developed a statistical analysis method named TreeHugger that considered both local and global structures of a gene ontology to assess the qualities of terms applied to annotate bioinformatics concepts. Our proposed methodology not only contributes to the design of an ontology evaluation method but also a novel approach of learning ontologies from folksonomies.

3. A Methodology of Learning Context-Sensitive Domain Ontologies from Folksonomies

The proposed methodology of learning lightweight, context-sensitive domain ontologies from folksonomies is depicted in Figure 1. The steps of the proposed methodology are as follows.

1. Text Preprocessing (Step 1): Social annotations of a specific domain (e.g., video games) extracted from folksonomy sites are cleaned up (e.g., removing low-frequency tags). Moreover, Web resources (e.g., Web pages) referred by tags are retrieved from respective Web sites for contextual aspects mining later on.
2. Tag Graph Generation (Step 2): Based on the statistical distributions of tags over Web resources and the measure of inclusion index (Salton 1986), the second step involves the construction of a tag subsumption graph to estimate domain-specific tag generality for taxonomic relation discovery.
3. Tag Graph Analysis (Step 3): The generality scores of tags captured in a tag subsumption graph are estimated according to a proximity prestige-based tag generality measure. Tag generality scoring facilitates the discovery and pruning of taxonomic relations among concepts. The computational details of Steps 2 and 3 will be illustrated in §4.2.
4. Concept Vector Mining (Step 4): Apart from link-based analysis of tags, the proposed methodology utilizes semantically rich concept vectors to address the social synonymy problem. In particular, a concept vector consists of a set of tags commonly used by a folksonomy community to describe a concept. For example, the concept "knowledge discovery" is represented by the concept vector $\langle (data, 0.519), (mining, 0.433), (database, 0.297), \ldots, \rangle$, and the concept "data mining" is represented by another vector $\langle (mining, 0.622), (data, 0.501), (ai, 0.119), \ldots, \rangle$. Accordingly, the semantic similarity between the concepts of "knowledge discovery" and "data mining" can be discovered by analyzing the underlying concept vectors. Concept vectors are mined from a

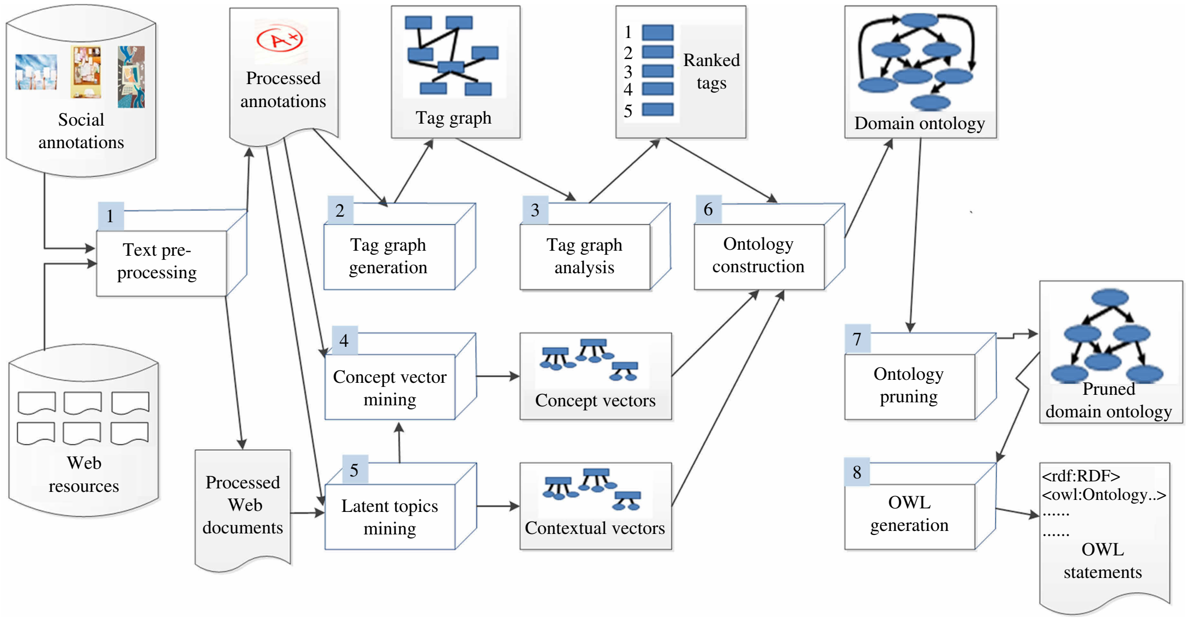


Figure 1 (Color online) A Methodology of Learning Context-Sensitive Domain Ontologies from Folksonomies

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

domain-specific corpus via the cognitively motivated basic-level concept mining algorithm that will be illustrated in §4.3 and §4.5, respectively.

5. Latent Topics Mining (Step 5): The proposed methodology leverages a LDA-based topic model to mine contextual aspects from a domain-specific corpus to address the social polysemy problem. For instance, by identifying the context "computing," the proposed system can infer that the tag "cloud" is about the concept "service." In contrast, the same tag is deduced to refer to the concept "atmospheric phenomena," given a mined context of "meteorology." The proposed LDA-based topic model is discussed in §4.4.

6. Domain Ontology Construction (Step 6): Given semantically rich representations of concepts and the contextual vectors representing the relevant "contexts" of ontology learning, §4.6 illustrates the proposed CSDOL algorithm, which can effectively infer taxonomic and nontaxonomic relations among domain concepts.

7. Domain Ontology Pruning (Step 7): Because of the ambiguity of natural languages, the main challenge of automatic ontology learning is how to prune noisy concepts and relations (Cimiano et al. 2005, Maedche and Staab 2004). Accordingly, the proposed methodology applies sound heuristics to remove noisy concepts and relations from an initially generated domain ontology. The computational details are explained in §4.6.

8. Visualization and Web Ontology Language Generation (Step 8): Finally, pruned domain ontologies are formally expressed using Web ontology language (http://www.w3.org/2004/OWL/) so that they are readily deployed to various Semantic Web applications. Visualization of learned domain ontologies is achieved via an extended version of the open-source graph display program named TouchGraph (http://sourceforge.net/projects/touchgraph/).

Some examples of the domain ontology generated by the proposed system are given in Appendix A of the online supplement (available as supplemental material at http://dx.doi.org/10.1287/ijoc.2015.0644).

4. The Computational Methods

4.1. Folksonomy and Ontology

Before illustrating the proposed algorithm of learning lightweight domain ontologies from folksonomies, the formal definitions of folksonomy and lightweight ontology are given as follows.

Definition 1 (Folksonomy). A folksonomy is a quadruple $F := \langle U, T, R, AN\rangle$, where U, T, and R are finite sets of users, tags, and resources, respectively; social annotations $AN \subseteq U \times T \times R$ are ternary relations among users, tags, and resources.

Definition 2 (Lightweight Ontology). A lightweight ontology is a sextuple $\mathrm{Ont} := \langle X, A, C, R_{CC}^{TAX}, R_{CC}^{NTAX}, U\rangle$, where X, A, C, and U stand for finite sets of objects, attributes, concepts, and users, respectively. The relations $R_{CC}^{TAX} \subseteq C \times C$ and $R_{CC}^{NTAX} \subseteq C \times C$ represent taxonomic and nontaxonomic relations among the set of concepts C perceived by the set of users U.

According to Definition 1, a folksonomy can be captured by a tripartite graph, with each annotation $(u, t, r) \in AN$ represented by a hyperedge. For a combined resource- and user-based mapping between a folksonomy and an ontology, each object $x_i \in X$ of an ontology is mapped to a Web resource $r_i \in R$. Each attribute $a_i \in A$ of x_i is characterized by a tag $t_k \in T$, and $U_k \subseteq U$ denotes the set of users who utilize the tag t_k to annotate the resource r_i. Moreover, a concept $c_i \in C$ of an ontology is an abstraction of a category of objects (e.g., Web resources) represented by the powerset 2^R. The proposed lightweight domain ontology is community-sensitive because the set of concepts C is induced from the perspective of the folksonomy community U. In this paper, objects, Web resources, and documents are considered interchangeable. The choice of a particular terminology depends on the context (e.g., objects for ontology, Web resources for folksonomy, and documents for topic modeling). Formally, an object is defined as follows.

Definition 3 (Object). An object $x_i \in X$ is represented by a vector $x_i = \langle (t_1, w(t_1)), (t_2, w(t_2)), \ldots, (t_n, w(t_n))\rangle$. Each tag $t_k \in T$ is created by some users $U_k \subseteq U$, who use t_k to annotate the object x_i. Each tag weight $w(t_k)$ quantifies the degree of importance or certainty of using the tag t_k to characterize the object x_i from the perspective of the folksonomy community $U_k \subseteq U$.

For a combined resource- and user-based mapping between a folksonomy and an ontology, both the tags and the users who create the tags are taken into account to characterize an object $x_i \in X$. In particular, each individual raw tag weight $rw(t_k)$ of the object x_i is defined by $rw(t_k) = U_{t_k}(x_i, t_k)/U(x_i)$, where $U_{t_k}(x_i, t_k)$ is a function that returns the number of users who apply tag t_k to annotate object x_i; the function $U(x_i)$ returns the total number of users who annotate the object x_i. After an initial object vector is formed, the raw tag weights of an object are subject to cosine normalization to derive the final tag weight $w(t_k) \in (0, 1]$; i.e., $w(t_k) = rw(t_k)/\sqrt{\sum_{l=1}^{n} rw(t_l)^2}$. The proposed formulation leverages the "collective intelligence" of a folksonomy community to define an object. This is one of the advantages of the proposed computational method and is different from some existing ontology learning methods that only consider the distributions of tags among Web resources

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

(Liu et al. 2010, Ponzetto and Strube 2007, Tang et al. 2009, Trabelsi et al. 2010, Zhou et al. 2007).

In the field of cognitive psychology, a concept is considered the abstraction of a category of objects that share some common attributes or features (Tanaka and Taylor 1991). Essentially, the proposed computational method induces a concept of a specific domain by identifying the set of common attributes (tags) applied to characterize a category of objects (Web resources) pertaining to that domain. The *extension* of a concept is defined by a set of objects categorized by the concept, and the *intension* of a concept is the common attributes (features) of the set of objects described by the concept (Cimiano et al. 2005, Lau et al. 2009). Each tag t_k is common to the set of objects $X_h \subseteq X$ categorized by a concept c_h. The weight $w(t_k) \in (0, 1]$ of a tag t_k for the concept c_h is the mean of the tag weights of t_k assigned to the set of objects categorized by the concept c_h. The intensional view of a concept is formally defined as follows.

Definition 4 (Concept). A concept $c_h \in C$ is an abstraction of some objects, and it is represented by a vector $c_h = \langle (t_1, w(t_1)), (t_2, w(t_2)), \ldots, (t_n, w(t_n))\rangle$. Each tag $t_k \in T$ is commonly used by some users $U_h \subseteq U$, who annotate a set of objects $X_h \subseteq X$ described by the concept c_h. Each tag weight $w(t_k)$ denotes the importance or certainty of using the tag t_k to characterize the concept c_h from the perspective of the folksonomy community $U_h \subseteq U$.

4.2. Domain-Specific Tag Subsumption Graph Generation and Analysis (Steps 2 and 3)

The objective of constructing a tag subsumption graph and analyzing the connectivity among tags is to induce a generality score for each tag to facilitate taxonomic relation discovery. Recently, the generality of information items (e.g., key words) has also been examined for improving domain-specific information retrieval (Yan et al. 2011). For the proposed methodology, we apply the generality scores of tags to verify the subsumption relations among concepts and prune redundant relations. For example, if a concept is characterized by many general tags (i.e., a general concept), it is more possible that this concept subsumes another concept, which is represented by many specific tags (i.e., a specific concept). A tag subsumption graph is built according to the Inclusion Index $\Pi(t_i, t_j)$ that measures how strongly a term t_i includes (i.e., subsumes) another term t_j (Salton 1986). It is grounded in the notion of conditional probability and is an asymmetric measure such that $\Pi(t_i, t_j) \neq \Pi(t_j, t_i)$ holds in general. When $\Pi(t_i, t_j) > \Pi(t_j, t_i)$ holds, the tag t_i is considered more general than the tag t_j. The Inclusion Index is defined by $\Pi(t_i, t_j) = \Pr(t_i \mid t_j) = n(t_i, t_j)/n(t_j)$, where $n(t_i, t_j)$ represents the number of Web resources annotated by both tags t_i and t_j, and $n(t_j)$ represents the number of Web resources annotated by tag t_j alone. The weight of an edge $w(e_{ij}) \in (0, 1]$ of a tag subsumption graph is defined by the net inclusion index score between tags t_i and t_j. Formally, a tag subsumption graph is defined as follows.

Definition 5 (Tag Subsumption Graph). A tag subsumption graph is a weighted directed acyclic graph $G = (V, E)$ that comprises finite sets of nodes V and edges E. Each node $v_i \in V$ represents a tag t_i. A directed edge $e_{ij} \in E$ indicates that tag t_i includes (subsumes) tag t_j. The weight of a directed edge $w(e_{ij}) \in (0, 1]$ represents the degree of inclusion from v_i to v_j.

For the proposed computational method, we estimate tag generality according to both a tag's local and global connectivities captured in a tag subsumption graph. More specifically, we extend the proximity prestige measure (Wasserman and Faust 1999) that has been widely used in social network analysis to estimate the domain-specific generality of a tag. In the context of tag generality computation, the intuition of the proposed method can be explained as follows: if a tag can reach its nonimmediate neighbors via many shortest paths (i.e., enjoying prestige reachness) in a tag subsumption graph, it implies that there is a greater potential for the tag to directly include (subsume) these nonimmediate neighboring tags. Accordingly, such a tag is considered a more general (common) tag. Formally, the reachability of a node in a tag subsumption graph is defined as follows.

Definition 6 (Directed Path). A directed path p_{ij} from a node v_i to another node v_j consists of a sequence of ordered nodes such as $(v_i, v_a), (v_a, v_b), \ldots, (v_d, v_j)$. Each pair of nodes along the path represents a directed edge from v_a to v_b, and all the pairs denote the same direction of the inclusion relation. The length of a directed path is defined by the number of unique nodes along the path p_{ij} minus one, that is, $\mathrm{len}(p_{ij}) = |\{v_i, v_a, \ldots, v_j\}| - 1$.

Definition 7 (Shortest Directed Path). Between any two nodes v_i and v_j, there exists a set of possible directed paths P_{ij}. The shortest path $p_{ij}^{\text{short}} \in P_{ij}$ between v_i and v_j holds the property $\exists p_{ij}^{\text{short}} \in P_{ij}\colon \forall p_{ij} \in P_{ij}, \mathrm{len}(p_{ij}) \geq \mathrm{len}(p_{ij}^{\text{short}})$.

Definition 8 (Directed Reachability). The directed reachability of a node v_i, denoted $\mathrm{reach}(v_i)$, is defined by the set of unique nodes along all the possible directed paths originated from v_i; i.e., $\mathrm{reach}(v_i) = \{v_j \in V\colon (v_i \neq v_j) \wedge (\mathrm{len}(p_{ij}) > 0)\}$.

The domain-specific generality measure $\mathrm{gen}_{tag}(v_i)$ is proposed to estimate the generality of each

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

node (tag) v_i of a tag subsumption graph, and it is defined as follows:

$$\text{gen}_{tag}(t_i) \approx \text{gen}_{tag}(v_i) = \frac{|\text{reach}(v_i)|/(|V|-1)}{(\sum_{v_j \in \text{reach}(v_i)} \text{len}(p_{ij}^{\text{short}}))/|\text{reach}(v_i)|}. \qquad (1)$$

The notation V represents the set of nodes of a tag subsumption graph; $\text{len}(p_{ij}^{\text{short}})$ represents the path length of the shortest directed path p_{ij}^{short}. Equation (1) represents the percentage of nodes reachable from a node v_i, divided by the average shortest path length to reach those nodes. The domain-specific generality score of a concept $c_h = \langle(t_1, w(t_1)), (t_2, w(t_2)), \ldots, (t_n, w(t_n))\rangle$ is derived based on the generality scores of the tags describing the concept. The concept generality score $\text{gen}(c_h)$ is defined by $\text{gen}(c_h) = (\sum_{t_i \in \text{dom}(c_h)} \text{gen}_{tag}(t_i) \times w(t_i))/|\text{dom}(c_h)|$, where $\text{dom}(c_h)$ represents the domain of the concept c_h and $w(t_i)$ is the tag weight of t_i for concept c_h.

4.3. Human Cognition-Based Concept Mining (Step 4)

To characterize the basic-level concepts most naturally recognized by humans, psychologists have developed a cognitively motivated measure named *category utility* (Gluck and Corter 1985). The results of many empirical tests show that the basic-level concepts perceived by humans often carry relatively high category utilities (Gluck and Corter 1985). From a computational point of view, category utility can be seen as a measure to quantify intracluster similarity and intercluster dissimilarity when humans apply domain knowledge to categorize objects. The details about the category utility measure are given in Appendix G of the online supplement. For the proposed concept mining algorithm, category utility is applied to extract prominent domain concepts from social annotations. In particular, features of objects are represented by tags. Since the proposed concept mining algorithm exploits the "collective intelligence" of a folksonomy community to induce the semantics of a concept, the importance of a tag as perceived by the community should be taken into account as well. Accordingly, we design the measure of weighted category utility $wcu(C, F)$ to mine basic-level concepts from the perspective of a folksonomy community:

$$wcu(C, F) = \frac{1}{n}\sum_{j=1}^{n} \Pr(c_j) \sum_{k=1}^{m} w(t_k)\big[\Pr(t_k \mid c_j)^2 - \Pr(t_k)^2\big], \qquad (2)$$

where C is a set of clusters (concepts) and F is a set of features involved in the categorization task. The notation $t_k \in F$ represents a tag (feature) that is used to characterize an object. The conditional probability $\Pr(t_k \mid c_j)$ captures the likelihood that an object of the cluster c_j is characterized by the tag t_k. The priori probability $\Pr(c_j)$ captures the likelihood that an object belongs to the cluster c_j, and $\Pr(t_k)$ is the priori probability that an object is characterized by the tag t_k. The terms n and m represent the total number of clusters and the total number of features, respectively. The normalization factor $1/n$ facilitates the comparison of the category utilities of partitions with varying sizes. $w(t_k)$ is the average weight of the tag t_k derived from the given social annotations. In particular, $w(t_k) = \sum_{l=1}^{|R_k|} w(t_k^l)/|R_k|$ is applied, where $w(t_k^l)$ is the weight of the tag t_k for a resource l and R_k is the set of resources annotated by t_k.

The proposed basic-level concept mining algorithm is depicted in Figure 2. First, every object (e.g., Web resource) $x_i \in X$ is taken as a concept $c_i \in C$. Second, the cosine similarity measure $sim(c_i, c_j)$ (Salton and McGill 1983) is applied to compute the similarity between each pair of concepts (c_i, c_j), and a concept similarity matrix S is built. Third, the most similar pair of concepts captured in the matrix S is identified and merged into a new concept c_{ij}. The newly combined concept contains all instances of the two constituent concepts and captures the common features of the constituent concepts; i.e., $feature(c_i) \cap feature(c_j)$. The feature extraction function $feature()$ returns the set of features pertaining to a concept. Then the similarity matrix S is updated by computing the similarities between the newly discovered concept and the remaining concepts captured by C. This iterative clustering procedure is repeated until all concepts are merged into a single one; i.e., $|C| = 1$. Finally, the level of clusters that produces the highest category utility is identified according to Equation (2). The clusters $CU[k]$ of the kth level that carries the highest category utility are extracted as the basic-level concepts.

4.4. Topic Modeling for Contextual Aspects Mining (Step 5)

According to previous studies in the field of cognitive psychology (Chun and Jiang 1998, Tanaka and Taylor 1991), *contexts* play an important role in human cognition. In particular, humans tend to evaluate the meaning of a concept with respect to a certain context during concept learning. In the field of computing, researchers have successfully applied context trees to model "contexts" for enhancing trend prediction and anomaly detection tasks (Brice et al. 2011). Accordingly, an effective computational method of learning ontologies from folksonomies should take into account contextual effects that influence the generation of social annotations. For instance, a "weather"-related resource represented by the object vector $x_i = \langle(cloud, 0.935),$

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

```
Algorithm BasicConceptMining($X, BC$)
Input: A set of objects $X$
Output: A set of basic-level concepts $BC$
Main Procedure:
$C \leftarrow X$, $F \leftarrow feature(X)$
$k \leftarrow 1$
$TC[\,] \leftarrow null$        // An array of concepts for each level
$CU[\,] \leftarrow null$        // An array of category utilities for each level
$BC \leftarrow \varnothing$, $S \leftarrow null$
Compute pairwise similarity over $C$ using a similarity metric $sim(\,)$
Construct the similarity matrix $S$ with $s[i,j] = sim(c_i, c_j)$ where $c_i \in C$, $c_j \in C$, $c_i \neq c_j$
while $|C| > 1$ do
    find $s[i,j] \in S$ with the highest similarity score
    $intent(c_{ij}) \leftarrow feature(c_i) \cap feature(c_j)$
    $extent(c_{ij}) \leftarrow c_i \cup c_j$
    $C \leftarrow C \cup c_{ij}$
    $C \leftarrow C - c_i$
    $C \leftarrow C - c_j$
    Compute pairwise similarity over $C$ and update $S$
    $TC[k] \leftarrow C$
    compute $CU[k] \leftarrow wcu(TC[k], F)$        // compute category utility at each level
    $k \leftarrow k + 1$
end
Find $CU[k]$ with the highest category utility
$BC \leftarrow TC[k]$
return $BC$
```

Figure 2 Basic-Level Concept Mining Algorithm

(*atmosphere*, 0.295), (*humidity*, 0.197)⟩ may be mistakenly judged as similar to another resource about "cloud service" represented by the object vector $x_j = \langle(cloud, 0.943), (service, 0.314), (internet, 0.105)\rangle$ if the cosine similarity of these resources is computed out of a context. By applying the standard cosine similarity measure, a relatively high cosine score of $sim(x_i, x_j) = 0.882/(1.016 \times 0.954) = 0.910$ is derived in this example. The consequence is that the proposed basic-level concept mining algorithm may mistakenly group these two distinct objects into a cluster (i.e., a potential basic-level concept) if it cannot take the appropriate "context" into account.

We propose the constructs of contextual vectors to represent a relevant "context." In particular, a contextual vector is represented by a vector of (term, value) pairs such as $cxt_h = \langle(t_1, w(t_1)), (t_2, w(t_2)), \ldots, (t_n, w(t_n))\rangle$ which satisfy $\sum_{k=1}^{n} w(t_k) = 1$. Intuitively, each term captured in a contextual vector indicates a significant contextual aspect that should be taken into account when basic-level concepts are mined from the corresponding domain. More specifically, we develop a LDA-based topic model (Blei et al. 2003, Rosen-Zvi et al. 2010) to dynamically extract contextual aspects from a domain-specific Web corpus to build contextual vectors. A Web corpus is constructed by retrieving the domain-specific Web documents and the corresponding social annotations. The intuition of topic modeling is that terms in documents are generated based on some latent topics that characterize these documents; such a generation process can be formalized according to a hierarchical Bayesian model (Blei et al. 2003).

For LDA-based topic modeling, each Web document $d \in D$ of an unlabeled training Web corpus D is characterized by a multinomial distribution θ, which is sampled from a Dirichlet prior with a hyperparameter α (Blei et al. 2003). The notation Z represents the set of latent topics characterizing D. A latent topic $z \in Z$ (i.e., some contextual aspects) is selected according to the multinomial distribution θ. Given the topic z, a term t is then generated according to the multinomial distribution ϕ, which is sampled from another Dirichlet priori with a hyperparameter β. Our ultimate goal is to infer the conditional probability $\Pr(t_k \mid z_i)$ that represents a latent topic (i.e., some contextual aspects). However, directly computing the multinomial distribution ϕ or θ is computationally very expensive. According to previous studies (Geman and Geman 1984), Gibbs sampling can be applied to efficiently compute the approximations $\bar\phi$ and $\bar\theta$ of the multinomial distributions ϕ and θ:

$$\bar\theta = \frac{C^{ZD}_{np} + \alpha}{\sum_{n' \in Z} C^{ZD}_{n'p} + |Z|\alpha}, \qquad \bar\phi = \frac{C^{VZ}_{mn} + \beta}{\sum_{m' \in V} C^{VZ}_{m'n} + |V|\beta}, \tag{3}$$

where C^{VZ}_{mn} is a count matrix that captures the number of times that term $t_k = m$ is assigned to the latent topic $z_i = n$, excluding the current word position. The notations m and n represent the indices of the count matrix C^{VZ}_{mn}. The set of vocabulary V is used to compose the

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

collection D, and the set of latent topics Z characterizes the collection D. The count matrix C^{ZD}_{np} records the number of times that the latent topic $z_i = n$ is assigned to a document $d_i = p$. The notations n and p represent the indices of the count matrix C^{ZD}_{np}. The computational complexity of Gibbs sampling is characterized by $O(I \cdot |Z| \cdot |D| \cdot |d_{\text{avg}}|)$, where I is the number of Gibbs iterations, $|Z|$ is the predefined number of latent topics, and $|d_{\text{avg}}|$ is the average length of a document d in the corpus D. The computational details of our LDA-based contextual aspects mining method are given in Appendix B of the online supplement.

4.5. Context-Sensitive Basic-Level Concept Mining (Step 4)

For context-sensitive basic-level concept mining, the weight $w(t_k)$ of a tag t_k should be adjusted when the weighted category utility (Equation (2)) is applied. More specifically, the tag weight should be adjusted according to the importance of the tag perceived in a particular context. Moreover, if a tag is not captured by any contextual vectors (i.e., a less significant term for that context), its tag weight should be discounted by a factor of ω. Let Cxt denote a set of contextual vectors representing a specific context, and $cxt_h \in Cxt$ is one of the contextual vectors. Moreover, W_{Cxt} represents the set of tag weights captured by all contextual vectors Cxt; $w(t_k^{cxt_h})$ denotes the weight of a tag t_k captured in the contextual vector cxt_h. The contextual weight $w(t_k^{Cxt})$ and the out-of-context weight $w(t_k^{\neg Cxt})$ of the tag t_k with respect to the set of contextual vectors Cxt are defined by

$$w(t_k^{Cxt}) = \max\bigl(\{w(t_k^{cxt_h}) \in W_{Cxt}\colon cxt_h \in Cxt\}\bigr), \quad (4)$$

$$w(t_k^{\neg Cxt}) = \min(W_{Cxt}) \times \omega, \quad (5)$$

where $\omega = 0.5$ is a predefined discount factor. The use of the maximum function in Equation (4) is similar to the application of a standard fuzzy disjunction operator to combine multiple plausible evidences for informational inference. The out-of-context weight $w(t_k^{\neg Cxt})$ of the tag t_k represents the discounted contextual weight if t_k is not captured by the set of contextual vectors Cxt. The context-sensitive weighted category utility $cswcu(C, F, Cxt)$ is then defined as follows:

$$cswcu(C, F, Cxt) = \frac{1}{n}\sum_{j=1}^{n} \Pr(c_j) \sum_{k=1}^{m}\bigl(w(t_k^{\text{Adj}}) \cdot w(t_k) \cdot [\Pr(t_k \mid c_j)^2 - \Pr(t_k)^2]\bigr), \quad (6)$$

$$w(t_k^{\text{Adj}}) = \begin{cases} w(t_k^{Cxt}) & \text{if } \exists\, cxt_h \in Cxt\colon\ t_k \in \text{dom}(Cxt_h), \\ w(t_k^{\neg Cxt}) & \text{otherwise.} \end{cases} \quad (7)$$

With reference to the basic-level concept mining algorithm depicted in Figure 2, the weighted category utility function (i.e., Equation (2)) should be replaced by the above context-sensitive weighted category utility function (i.e., Equation (6)). Moreover, the cosine similarity between two objects should be adjusted with respect to the set of relevant contextual vectors Cxt as follows:

$$sim(x_i, x_j, Cxt) = \frac{\sum_{k=1}^{n} w(t_k^{\text{Adj}}) \times w(t_k^i) \times w(t_k^j)}{\sqrt{\sum_{k=1}^{n} w(t_k^i)^2} \times \sqrt{\sum_{k=1}^{n} w(t_k^j)^2}}, \quad (8)$$

where $w(t_k^{\text{Adj}}) = w(t_k^{Cxt})$ if the tag t_k is captured by the set of contextual vectors Cxt. The term $w(t_k^i)$ represents the tag weight of the tag t_k pertaining to the object x_i. With reference to the previous example of computing the similarity between the "weather" object x_i and the "cloud service" object x_j, the weighted cosine similarity function (Equation (8)) returns a relatively low value; i.e., $sim(x_i, x_j, Cxt) = 0.069/(1.016 \times 0.954) = 0.071$ with respect to the context "meteorology" represented by one contextual vector $cxt_{meteorology} = \langle(atmosphere, 0.821), (humidity, 0.101), (cloud, 0.078)\rangle$. This simple example shows that the result of basic-level concepts mining can be largely influenced by the underlying context. In fact, Equations (6) and (8) are designed to simulate how humans apply domain knowledge to categorize objects given a specific context. As a whole, extracting relevant contextual aspects and applying this knowledge to guide domain concept mining (e.g., measuring similarity among objects and computing category utility) is essential for effective learning of context-sensitive domain ontologies.

4.6. Context-Sensitive Domain Ontology Construction and Pruning (Steps 6 and 7)

For domain ontology construction, we employ a subsumption-based approach to identify taxonomic and nontaxonomic relations among domain concepts (Lau et al. 2009, Sanderson and Croft 1999). According to formal concept analysis, the intensions of concepts form the basis to evaluate the subsumption relations among these concepts (Cimiano et al. 2005, Tho et al. 2006). Given the *intensions* of two concepts c_i and c_j, if all the attributes of c_i also belong to c_j, i.e., $\{t_1^i, t_2^i, \ldots, t_n^i\} \subset \{t_1^j, t_2^j, \ldots, t_n^j\}$, the concept c_i is said to subsume the concept c_j (Cimiano et al. 2005). Because of the ambiguity of natural languages and the imperfect knowledge representation in computers, there is always a fuzziness (uncertainty) about the subsumption relations among concepts (Lau et al. 2009, Tho et al. 2006). Accordingly, the subsumption relation between two concepts (i.e., the inclusion relation between two sets of attributes) is better expressed as a degree of subsumption rather than a strict binary subsumption relation.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

In this paper, we propose a novel context-sensitive subsumption function that takes into account the inclusion relation between the sets of tags (attributes) describing two concepts c_i and c_j and the average generality of the tags pertaining to these concepts to estimate the degree of subsumption between the concepts c_i and c_j. To evaluate the inclusion relation between two concepts, a contextually weighted inclusion function is applied to measure the overlapping intensions of these concepts. The basic intuition is that an inclusion relation is strong with respect to a specific context represented by *Cxt* if the tags involved in the evaluation of the inclusion relation are considered important attributes in that context. For example, the tag "atmosphere" is considered a more important attribute than the tag "software" to characterize the context "meteorology." The context-sensitive subsumption function $subsume(c_i, c_j, Cxt)$ for a pair of concepts c_i and c_j with respect to the context *Cxt* is defined by

$$include(c_i, c_j, Cxt) = \frac{\sum_{t_k \in \{c_i \cap c_j\}} [w(t_k) \times w(t_k^{\mathrm{Adj}})]}{\sum_{t_l \in c_j} [w(t_l) \times w(t_l^{\neg Cxt})]}, \quad (9)$$

$$generalize(c_i, c_j) = \begin{cases} \mathrm{gen}(c_i) - \mathrm{gen}(c_j) & \text{if } \mathrm{gen}(c_i) > \mathrm{gen}(c_j), \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

$$subsume(c_i, c_j, Cxt) = \gamma \cdot include(c_i, c_j, Cxt) + (1-\gamma) \cdot generalize(c_i, c_j), \quad (11)$$

where the $include(c_i, c_j, Cxt)$ function measures the degree of inclusion from concept c_i to concept c_j with respect to the context represented by *Cxt*.

In contrast, the function $generalize(c_i, c_j)$ measures the relative generality between two concepts c_i and c_j. The term $\mathrm{gen}(c_i)$ is the generality score of the concept c_i as defined in §4.2. The function $generalize(c_i, c_j)$ is domain sensitive because a tag subsumption graph that induces the generality scores of concepts is constructed based on a domain-specific corpus. Finally, the degree of subsumption from c_i (superconcept) to c_j (subconcept) is a weighted sum of the functions $include(c_i, c_j, Cxt)$ and $generalize(c_i, c_j)$. The intuition of equation (Equation (11)) is that it follows the format of the standard convex function that has been successfully applied to a variety of applications; it represents a novel way of combining link- and content-based approaches for the discovery of subsumption relations among concepts. For the experiments reported in this paper, we assigned equal weight to the inclusion function and the generality function (i.e., $\gamma = 0.5$) according to the initial parameter tuning conducted based on the "software" domain.

The algorithm for context-sensitive domain ontology construction and pruning is depicted in Figure 3. First, the subsumption scores among the set of basic-level concepts mined from a given domain-specific corpus are computed. If both $subsume(c_i, c_j, Cxt) > th_{\mathrm{sub}}$ and $subsume(c_j, c_i, Cxt) > th_{\mathrm{sub}}$ are established, a potential semantic relation between c_i and c_j exists. The subsumption threshold th_{sub} is used to filter noisy relations among concepts. In addition, if $[subsume(c_i, c_j, Cxt) - subsume(c_j, c_i, Cxt)] > th_{\mathrm{gap}}$ is established, a taxonomic relation between the superconcept c_i and the subconcept c_j is identified. The reason is that c_i has a higher tendency to subsume c_j with respect to the specific context. The relation threshold th_{gap} is applied to distinguish whether the relation between two concepts is a taxonomic relation or a similarity (association) relation. Both th_{sub} and th_{gap} are empirically established based on a training corpus. After the assertions of all the relationships among concepts, multiple concept hierarchies may be generated. All these concept hierarchies are attached to a root node that represents the most general concept of a specific domain (e.g., software, hardware, business, etc.).

The initial domain ontology may be noisy because redundant and inconsistent taxonomic relations exist. For instance, if

$$[subsume(c_i, c_j, Cxt) - subsume(c_j, c_i, Cxt)] > th_{\mathrm{gap}},$$
$$[subsume(c_j, c_k, Cxt) - subsume(c_k, c_j, Cxt)] > th_{\mathrm{gap}},$$

and

$$[subsume(c_k, c_i, Cxt) - subsume(c_i, c_k, Cxt)] > th_{\mathrm{gap}},$$

are all established, the induced subgraph of the ontology contains redundant and inconsistent subsumption relations (i.e., $c_i \to c_j \to c_k \to c_i$). Our ontology pruning procedure evaluates the minimal degree of subsumption along the path p: $c_i \to c_j \to c_k$, where $\to$ indicates a "subsumption" relation. If the minimal subsumption degree along the path p is greater than $subsume(c_k, c_i, Cxt)$, the relation $c_k \to c_i$ should be removed because it is a weaker inconsistent subsumption relation. If

$$[subsume(c_i, c_j, Cxt) - subsume(c_j, c_i, Cxt)] > th_{\mathrm{gap}},$$
$$[subsume(c_j, c_k, Cxt) - subsume(c_k, c_j, Cxt)] > th_{\mathrm{gap}},$$

and

$$[subsume(c_i, c_k, Cxt) - subsume(c_k, c_i, Cxt)] > th_{\mathrm{gap}},$$

are all established (i.e., $c_i \to c_j \to c_k$ and $c_i \to c_k$), the relation $c_i \to c_k$ is considered redundant because the path p: $c_i \to c_j \to c_k$ implies the relation $c_i \to c_k$ via a transitivity closure.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

```
Algorithm OntLearn($BC, root, Cxt, \gamma, th_{\text{sub}}, th_{\text{gap}}, \text{Max}_{\text{sub}}, \text{Max}_{\text{sib}}$)
Inputs: A set of basic-level concepts $BC$, a default domain-specific loot node $root$, a relevant context $Cxt$, the weight
        factor for subsumption score computation $\gamma$, a minimum subsumption score $th_{\text{sub}}$, a relation threshold
        between a pair of concepts $th_{\text{gap}}$, maximal number of subconcepts of each node $\text{Max}_{\text{sub}}$, maximal number of
        siblings of each node $\text{Max}_{\text{sib}}$
Output: a lightweight domain ontology $\text{Ont} := \langle X, A, C, R_{CC}^{TAX}, R_{CC}^{NTAX} \rangle$
Main Procedure:
Ont $\leftarrow$ root
// compute subsumption relations among each pair of concepts
for each pair of distinct concepts $c_i \in BC$, $c_j \in BC$, $c_i \neq c_j$ do
    compute $subsume(c_i, c_j, Cxt) = \gamma \cdot include(c_i, c_j, Cxt) + (1-\gamma) \cdot generalize(c_i, c_j)$
    compute $subsume(c_j, c_i, Cxt) = \gamma \cdot include(c_j, c_i, Cxt) + (1-\gamma) \cdot generalize(c_j, c_i)$
    if $subsume(c_i, c_j, Cxt) > th_{\text{sub}}$ and $subsume(c_j, c_i, Cxt) > th_{\text{sub}}$
        $X \leftarrow X \cup objects(c_i, c_j)$        // retrieve the objects characterized by $c_i$, $c_j$
        $A \leftarrow A \cup attributes(c_i, c_j)$        // retrieve the attributes of $c_i$, $c_j$
        $C \leftarrow C \cup c_i \cup c_j$        // save concepts $c_i$, $c_j$
        if $[subsume(c_i, c_j, Cxt) - subsume(c_j, c_i, Cxt)] > th_{\text{gap}}$        // a subsumption relation
            $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} \cup (c_i, c_j)$
        else if $[subsume(c_j, c_i, Cxt) - subsume(c_i, c_j, Cxt)] > th_{\text{gap}}$        // a subsumed-by relation
            $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} \cup (c_j, c_i)$
        else
            $R_{CC}^{NTAX} \leftarrow R_{CC}^{NTAX} \cup (c_i, c_j)$        // nontaxonomic association relation
        end
    end
end
end
// ontology pruning
if $\exists (c_i \rightarrow c_x, \ldots, c_y \rightarrow c_j)$ and $(c_i \rightarrow c_j)$        // redundant relation
    $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} - (c_i, c_j)$
end
if $\exists (c_i \rightarrow c_x, \ldots, c_y \rightarrow c_j)$ and $(c_j \rightarrow c_i)$        // inconsistent relation
    if $\min(\{subsume(c_i, c_x, Cxt), \ldots, subsume(c_y, c_j, Cxt)\}) > subsume(c_j, c_i, Cxt)$
        $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} - (c_j, c_i)$        // remove the weaker path $(c_j \rightarrow c_i)$
    else if $\min(\{subsume(c_i, c_x, Cxt), \ldots, subsume(c_y, c_j, Cxt)\}) = subsume(c_j, c_i, Cxt)$
        $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} - (c_j, c_i)$        // remove the shorter, less informative path
    else
        $R_{CC}^{TAX} \leftarrow R_{CC}^{TAX} - \{(c_i, c_x), \ldots (c_y, c_j)\}$        // $(c_j \rightarrow c_i)$ is stronger
    end
    end
end
for each $c_i \in C$ do
    prune $R_{CC}^{TAX}$ and $R_{CC}^{NTAX}$ according to $\text{Max}_{\text{sub}}$ and $\text{Max}_{\text{sib}}$
end
return Ont
```

Figure 3 **Context-Sensitive Domain Ontology Construction and Pruning**

For each concept c_i, there are a maximum number of subconcepts Max_{sub} and a maximum number of siblings Max_{sib} connected via association relations. These thresholds are applied to further prune the relatively weak relations associated with each concept. When compared to other existing ontology learning algorithms (Heymann and Garcia-Molina 2006, Liu et al. 2010, Trabelsi et al. 2010, Zhou et al. 2007), our proposed algorithm allows a concept to have multiple superconcepts. As a result, richer semantics can be captured by the domain ontologies generated by our system. The structural characteristics of our domain ontologies are more akin to the real-world ontologies such as Open Directory Project and WordNet, which support multiple inheritances. The computational complexity of the domain ontology construction algorithm is characterized by $O(|BC|^2)$, where $|BC|$ is the cardinality of the set of basic-level concepts.

5. System Evaluation

5.1. General Experimental Procedure and Data Set

We evaluated the proposed methodology from three perspectives: structural, semantic, and comparative assessments (Zavitsanos et al. 2011, Zouaq and Nkambou 2009). Structural assessment aims to objectively evaluate the structural properties such as the depth and the number of concept nodes of an ontology. Semantic evaluation involves the assessment of the semantic contents of an ontology generated by a system with respect to a gold standard ontology developed by

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

humans. Finally, comparative assessment involves the evaluation of the domain ontologies generated by the proposed system and those produced by other existing systems. We adopted the widely used concept hierarchy generated by the Open Directory Project as the gold standard (Lau et al. 2009, Liu et al. 2010, Tang et al. 2009, Zhou et al. 2007). Our experiments were performed based on ten randomly chosen domains and three different sources of social annotations. For the domain of "movie," our crawler program retrieved the social annotations from imdb.com. In addition, our crawler program downloaded the data from citeulike.org for the domain of "artificial intelligence." For the remaining domains such as "business," "video game," "software," "hardware," "medicine," "clothing," "travel," and "cooking," we crawled the social annotations from delicious.com. All the social annotations and Web resources were crawled between February and December 2011. After removing tags and resources with low frequencies, the actual evaluation data set contains 128,506 users, 14,303 Web resources with valid URLs, 4,689 tags, and 12,766,436 annotations. The details of our evaluation data set are provided in Appendix C of the online supplement.

5.2. Evaluation Metrics

Similar to previous studies, both the lexical layer (e.g., concepts) and the taxonomic layer (e.g., structures) of the domain ontologies generated by a system were evaluated (Liu et al. 2010, Zavitsanos et al. 2011, Zouaq and Nkambou 2009). The evaluation metrics of lexical (taxonomic) precision, lexical recall, and lexical F-measure were applied to our experiments. For the lexical layer, we assessed the quality of the domain ontologies generated by a system based on how well these ontologies captured the concepts encoded in the gold standard ontologies. Lexical precision, lexical recall, and lexical F-measure are the same as the common evaluation metrics such as precision, recall, and F-measure widely used in information retrieval research (Salton and McGill 1983). Given a system-generated domain ontology Ont_{sys} and a gold standard ontology Ont_{gold}, taxonomic precision, taxonomic recall, and taxonomic F-measure are defined based on the notion of common semantic cotopy (Cimiano et al. 2005, Liu et al. 2010). We adopted $\beta = 1$ to weight precision and recall equally for all the experiments reported in this paper. The notation $\rightarrow_{\mathrm{Ont}}$ denotes a subsumption relation between a pair of concepts with respect to a given ontology Ont. The formal definitions of semantic cotopy, taxonomic precision, taxonomic recall, and taxonomic F-measure are given as follows.

Definition 9 (Semantic Cotopy). Given an ontology Ont, the semantic cotopy $\mathrm{SC}(\mathrm{Ont}, c_h)$ of a concept $c_h \in C$ is the set of all its direct and indirect super- and subconcepts, including the concept itself; i.e.,

$$\mathrm{SC}(\mathrm{Ont}, c_h) = \big\{\forall c_i \in C\colon (c_h \rightarrow_{\mathrm{Ont}} c_i) \vee (c_i \rightarrow_{\mathrm{Ont}} c_h)\big\} \cup c_h.$$

Definition 10 (Common Semantic Cotopy). Given two ontologies Ont_{sys} and Ont_{gold}, the common semantic cotopy of a concept $c_h \in C_{sys}$ is the set of all common super- and subconcepts of c_h; i.e., $\mathrm{CSC}(\mathrm{Ont}_{sys}, \mathrm{Ont}_{gold}, c_h) = \{\forall c_i \in \{C_{sys} \cap C_{gold}\}\colon (c_h \rightarrow_{sys} c_i) \vee (c_i \rightarrow_{sys} c_h)\}$.

$$Precision_{tax} = \frac{\sum_{c_h \in \{\mathrm{Ont}_{sys} \cap \mathrm{Ont}_{gold}\}} \mathrm{CSC}(\mathrm{Ont}_{sys}, \mathrm{Ont}_{gold}, c_h)}{|C_{sys} \cap C_{gold}|}, \qquad (12)$$

$$Recall_{tax} = \frac{\sum_{c_h \in \{\mathrm{Ont}_{sys} \cap \mathrm{Ont}_{gold}\}} \mathrm{CSC}(\mathrm{Ont}_{gold}, \mathrm{Ont}_{sys}, c_h)}{|C_{sys} \cap C_{gold}|}, \qquad (13)$$

$$F_{tax} = \frac{(1+\beta^2) Precision_{tax} \times Recall_{tax}}{\beta^2\, Precision_{tax} + Recall_{tax}}. \qquad (14)$$

5.3. Testing the Internal Validity of the Proposed Methodology

Our first validity test examined whether the proposed CSDOL method was more effective than its context-free counterpart. For context-free ontology learning, the standard cosine similarity measure (Salton and McGill 1983) and the context-free weighted category utility Equation (2) were applied to extract domain concepts. Moreover, only the out-of-context weight $w(t_k^{\neg Cxt})$ was applied to Equation (9) to compute the inclusion scores of concepts. In contrast, context-sensitive ontology learning invoked the context-aware cosine similarity measure Equation (8) and the context-aware weighted category utility function Equation (6) to mine domain concepts with respect to the relevant context represented by a set of contextual vectors Cxt.

The context-sensitive ontology learning system (CSDOL) invoked the LDA-based topic model to extract contextual vectors from each corpus before a domain ontology was learned. We ran two independent Gibbs sampling chains for a maximum of 500 iterations each because it is a common practice to invoke multiple Gibbs sampling chains to ensure proper convergence of the sampling process (Tang et al. 2009). For all experiments reported in this paper, $th_{\mathrm{sub}} = 0.05$, $th_{\mathrm{gap}} = 0.06$, $\gamma = 0.5$, and $\mathrm{Max}_{\mathrm{sub}} = \mathrm{Max}_{\mathrm{sib}} = 16$ were applied. All system parameter values were first empirically established based on the corpus of "software." Based on the common cross-validation approach (Yan et al. 2011), the established parameter values were then applied to other test domains. Although the parameter values established using such an approach may not be the

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

Table 2 Comparative Lexical Performance of Context-Sensitive vs. Context-Free Ontology Learning

	Lexical recall (%)		Lexical precision (%)		Lexical F-measure (%)	
Domain	Context-sensitive	Context-free	Context-sensitive	Context-free	Context-sensitive	Context-free
Movie	23.84	21.82	88.06	60.34	37.52	32.05
Business	24.24	22.83	87.59	60.11	37.97	33.09
Video game	22.22	20.61	82.71	54.84	35.03	29.96
Software	26.46	23.43	90.34	60.73	40.94	33.82
Hardware	27.68	25.45	92.57	62.38	42.61	36.15
Medicine	27.07	25.66	89.93	63.18	41.61	36.49
Clothing	23.03	22.83	86.36	59.16	36.36	32.94
Travel	24.04	22.42	88.15	58.42	37.78	32.41
Cooking	25.05	24.65	88.57	60.70	39.06	35.06
AI	16.57	15.15	65.08	54.35	26.41	23.70
Average	24.02	22.48	85.94	59.42	**37.53**	32.57

global optimum, a more sophisticated parameter tuning method (e.g., using genetic algorithms) will only further improve these results. We will leave the search for optimal system parameter values as part of our future research. We extracted the meta field, the title field, all heading fields, and all highlighted fields from each Web document of an evaluation domain to build a domain-specific corpus. Standard text preprocessing methods such as stop word removal, case transformation, and word stemming were applied to each Web corpus (Salton and McGill 1983).

The comparative lexical performance of the context-sensitive and the context-free systems is shown in Table 2. The context-sensitive system outperforms its context-free counterpart in terms of lexical F-score for all test domains because of its capability of alleviating the social polysemy problem. The average lexical F-scores achieved by the context-sensitive and the context-free ontology learning systems are 37.53% and 32.57%, respectively. The former system achieves a significant improvement of 15.24% in terms of average lexical F-score ($t(9) = 8.4$, $p < 0.01$ with paired one-tail t-test). The differences of the lexical F-scores achieved by these two systems are large for most domains. For the "AI" domain, neither system performed well. The reason might be that our "AI" corpus was constructed based on a set of academic publications, whereas the Open Directory Project directory mostly refers to commercial Web pages. As a result, both systems could only learn a few domain concepts captured by the Open Directory Project ontology.

The average taxonomic F-scores achieved by the context-sensitive and the context-free ontology learning systems with respect to various threshold values of th_{sub} and th_{gap} are plotted in Figures 4 and 5, respectively. A careful inspection of our experimental results showed that if th_{sub} was too small, some trivial relations would be classified as taxonomic relations by both systems. In contrast, many valid semantic relations were filtered out if th_{sub} was set to a relatively large value (e.g., $th_{\text{sub}} \geq 0.2$). Similarly, if th_{gap} was too small (e.g., $th_{\text{gap}} = 0.01$), many nontaxonomic relations were mistakenly classified as taxonomic relations. In contrast, if th_{gap} was too large (e.g., $th_{\text{gap}} \geq 0.09$), many taxonomic relations were

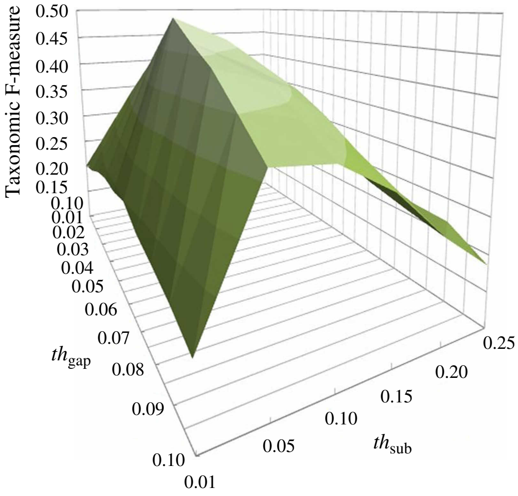


Figure 4 (Color online) Taxonomic F-Score Achieved by Context-Sensitive Ontology Learning System

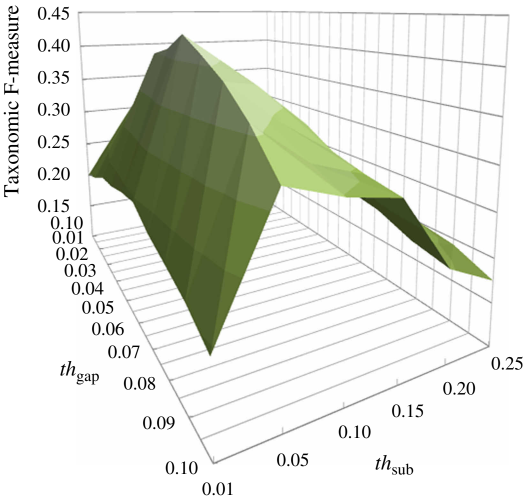


Figure 5 (Color online) Taxonomic F-Score Achieved by Context-Free Ontology Learning System

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

incorrectly labeled as nontaxonomic. Both the context-sensitive and the context-free ontology learning systems demonstrated similar performance fluctuation patterns with respect to various combinations of the th_{sub} and the th_{gap} parameter values. Nevertheless, the average taxonomic F-score achieved by the context-sensitive ontology learning system (e.g., 49.31%) is consistently higher than that achieved by its context-free counterpart (e.g., 42.85%). The detailed taxonomic F-scores achieved by both systems are provided in Appendix D of the online supplement.

Figure 6 shows a segment of the ontology learned by the context-sensitive system for the domain "software." For readability reason, the figure only depicts the drilling down of one concept node at each level of the ontology. Through this experiment, we can assert that the proposed CSDOL method is more effective than its context-free counterpart. We also repeated the same experiment using the full text of each Web document. However, we did not observe significant differences in the experimental results. Accordingly, each domain-specific corpus was constructed based on the textual contents of the key fields for the remaining experiments reported in this paper. For the second validity test, we aimed to test the effectiveness of a hybrid link- and content-based approach (i.e., tag subsumption graph and concept vector) compared to a purely content-based approach (i.e., concept vector alone) for domain ontology learning. The experimental results are provided in Appendix E of the online supplement. In addition, a user-based evaluation of the cognitively motivated basic-level concept learning method compared to the hierarchical K-means clustering method (Munir et al. 2012) will be discussed in §5.5.

5.4. Comparative Structural and Semantic Evaluation

For a comparative system evaluation, we employed two well-known folksonomy-based ontology learning methods: closeness centrality-based tag graph (CCTG) (Heymann and Garcia-Molina 2006) and tag association rule mining (TARM) (Schmitz 2006) because comprehensive documentations of these methods are available for baseline system development. In addition, the hierarchical K-means clustering method (Munir et al. 2012) (HK-means) was adopted to develop the third baseline system. For HK-means clustering, the baseline system referred to the Open Directory Project directory to estimate the initial number of clusters K. After each clustering step, if there were more than K objects left in a cluster, the system repeatedly applied the HK-means algorithm to further cluster the objects until leaf nodes were formed.

For the first comparative study, we examined the structural properties (e.g., ontology depth) of the domain ontologies generated by CSDOL and those produced by other baseline systems. Figures 7 (the software domain) and 8 (the business domain) show typical shapes of domain ontologies produced by the experimental and the baseline systems. The proposed CSDOL system generated a deeper ontology than

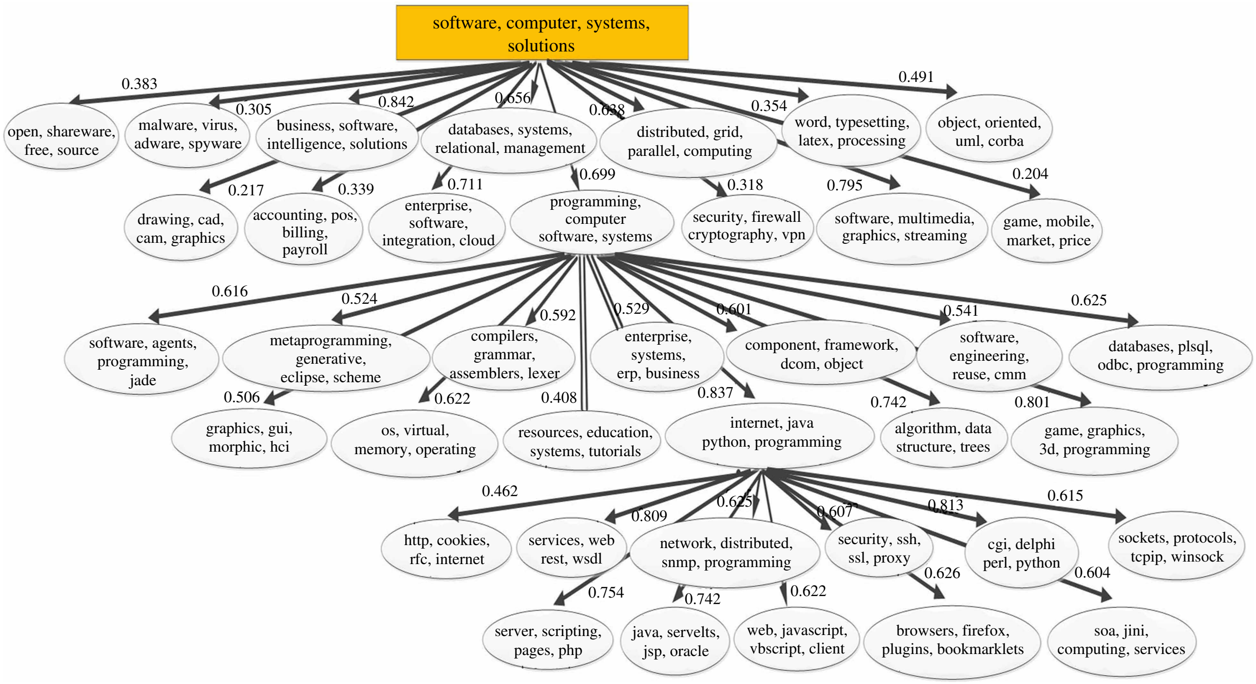


Figure 6 (Color online) A Sample Segment of the Ontology Learned by the CSDOL System

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

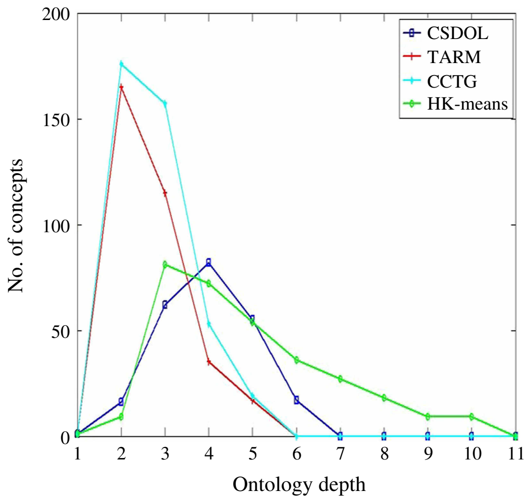


Figure 7 (Color online) Structural Properties of the Ontologies for the "Software" Domain

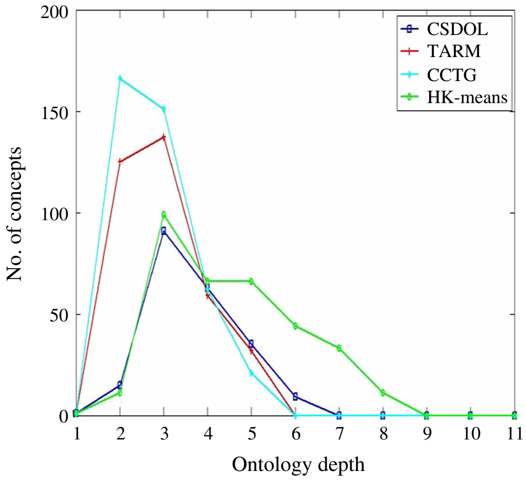


Figure 8 (Color online) Structural Properties of the Ontologies for the "Business" Domain

both the CCTG and the TARM systems did. Both the CCTG and the TARM baseline systems generated a large number of concepts (including noisy concepts) at the first three levels of an ontology because they employed a direct one-to-one mapping approach between tags and concepts. The HK-means system generated a much deeper ontology structure than humans did (e.g., the Open Directory Project ontology) because it employed a context-free, recursive approach to blindly divide a cluster into subclusters until the predefined limit K was satisfied. The ontological structures generated by the CSDOL system seem able to capture the typical concepts and semantic relations as exhibited in a variety of domains.

Using the Open Directory Project as the standard, we applied both lexical and taxonomic F-measures to assess the domain ontologies generated by the CSDOL system and those generated by other baseline systems. Table 3 depicts the lexical performance achieved by these systems. The average lexical F-score achieved by the CSDOL system is 37.53%, which is significantly higher than those obtained by other baseline systems according to paired one-tail t-tests. The CSDOL system significantly outperforms the best baseline system TARM by 7.42% ($t(9) = 8.4, p < 0.01$). According to postexperiment inspection, both the CCTG and the TARM systems generated many concepts and relations because they adopted a one-to-one mapping approach between tags and concepts. As a result, both systems achieved a recall comparable to that of the CSDOL system. However, the precision of these baseline systems is much lower than that of the CSDOL system.

For example, the TARM system simply maps the tag "business objects" to a concept called "business objects," whereas "business objects" is only taken as an instance of the concept "decision support software" by the Open Directory Project ontology. In contrast, the CSDOL system utilizes a cognitively motivated basic concept-mining algorithm that learns semantically rich representations of domain concepts from folksonomies. For instance, the

Table 3 Comparative Lexical Performance of Different Ontology Learning Systems

	Lexical recall (%)				Lexical precision (%)				Lexical F-measure (%)			
Domain	CSDOL	CCTG	TARM	HK-means	CSDOL	CCTG	TARM	HK-means	CSDOL	CCTG	TARM	HK-means
Movie	23.84	23.23	23.43	17.37	88.06	60.53	64.09	53.42	37.52	33.58	34.32	26.22
Business	24.24	23.64	23.84	17.17	87.59	61.58	62.77	53.80	37.97	34.16	34.55	26.03
Video game	22.22	22.22	22.02	15.35	82.71	54.73	56.19	45.51	35.03	31.61	31.64	22.96
Software	26.46	26.06	26.87	19.60	90.34	64.18	65.52	57.40	40.94	37.07	38.11	29.22
Hardware	27.68	27.27	28.08	20.40	92.57	64.90	66.83	58.05	42.61	38.41	39.54	30.19
Medicine	27.07	26.67	27.88	20.20	89.93	65.02	67.32	58.48	41.61	37.82	39.43	30.03
Clothing	23.03	22.02	23.23	19.19	86.36	59.24	61.50	54.91	36.36	32.11	33.72	28.44
Travel	24.04	23.84	24.04	19.80	88.15	61.46	61.66	56.98	37.78	34.35	34.59	29.39
Cooking	25.05	25.05	26.26	20.81	88.57	62.94	64.68	58.52	39.06	35.84	37.36	30.70
AI	16.57	15.15	16.57	14.34	65.08	59.06	61.65	54.62	26.41	24.12	26.11	22.72
Average	24.02	23.52	24.22	18.42	85.94	61.36	63.22	55.17	**37.53**	33.91	34.94	27.59

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

Table 4 **Comparative Taxonomic Performance of Different Ontology Learning Systems**

	Taxonomic recall (%)				Taxonomic precision (%)				Taxonomic F-measure (%)			
Domain	CSDOL	CCTG	TARM	HK-means	CSDOL	CCTG	TARM	HK-means	CSDOL	CCTG	TARM	HK-means
Movie	30.18	31.53	30.63	21.40	84.81	67.96	76.84	62.09	44.52	43.08	43.80	31.83
Business	30.86	32.66	30.41	22.75	88.39	69.05	77.59	65.58	45.74	44.34	43.69	33.78
Video game	27.70	22.75	22.75	18.24	78.34	58.72	59.76	53.29	40.93	32.79	32.95	27.18
Software	40.32	35.81	32.88	23.20	90.86	71.30	78.92	65.19	55.85	47.68	46.42	34.22
Hardware	40.99	34.91	31.76	23.65	91.92	73.81	82.46	66.04	56.70	47.40	45.85	34.83
Medicine	38.51	34.69	34.68	23.42	90.96	73.33	82.80	66.24	54.11	47.09	48.89	34.61
Clothing	32.66	32.88	31.53	21.62	87.88	72.28	76.50	61.94	47.62	45.20	44.66	32.05
Travel	40.54	30.86	30.86	20.72	91.84	67.82	76.97	62.59	56.25	42.41	44.05	31.13
Cooking	38.06	30.41	31.76	21.40	88.95	68.88	80.11	64.63	53.31	42.19	45.48	32.15
AI	25.23	23.65	24.77	22.30	77.24	72.92	75.86	69.72	38.03	35.71	37.35	33.79
Average	34.50	31.01	30.20	21.87	87.12	69.61	76.78	63.73	**49.31**	42.79	43.32	32.56

CSDOL system discovers the basic-level concept "decision support software" represented by the concept vector: $\langle(\mathit{intelligence}, 0.831), (\mathit{software}, 0.805), (\mathit{business}, 0.729), (\mathit{decision}, 0.617), (\mathit{support}, 0.604), (\mathit{objects}, 0.563), \ldots\rangle$. This is indeed an example of how the CSDOL system alleviates the social synonymy problem. As a result, a better match was found between the concepts discovered by the CSDOL system and those captured by the Open Directory Project ontology. The HK-means system achieves the worst lexical performance because it cannot utilize domain-specific knowledge to mine concepts.

For the taxonomic assessment, the comparative performance of various systems is shown in Table 4. The taxonomic F-score achieved by the CSDOL system is significantly higher than those obtained by the baseline systems according to paired one-tail t-tests. For example, the CSDOL significantly outperforms the best baseline system TARM by 13.83% in terms of taxonomic F-score ($t(9) = 4.4, p < 0.01$). Similar to the comparative evaluation at the lexical level, all systems perform relatively poorly for the "AI" and the "video game" domains. For the "AI" domain, Web resources of our corpus and that of the Open Directory Project ontology are quite different. The "video game" domain is a relatively noisy domain that mainly consists of the names of different games.

5.5. User-Based Semantic Evaluation

For the user-based evaluation, we invited 20 domain experts who were postgraduate students with at least four years' experience for the respective domains to assess the domain concepts and the semantic relations discovered by the proposed system and the HK-means baseline system, respectively. For each domain, two experts were asked to examine the concepts and semantic relationships captured at the top three levels of the corresponding domain ontologies. We focus on the top three levels because the concept hierarchy captured at the top of an ontology plays a more important role in determining the overall quality of the ontology than the bottom elements do. Only if both experts agreed that a concept (semantic relation) was correct would it be considered a correct concept (semantic relation) discovered by a system. The average interrater agreement as measured by Cohen's kappa is $\mathcal{K} = 0.77$, which indicates a relatively consistent and reliable expert judgment. Since it is difficult for the experts to articulate the complete "ground truth" for each domain, we can only evaluate the lexical precision of the respective systems. For this user-based evaluation, the measure of lexical precision was applied to evaluate both the domain concepts and the semantic relations among concepts.

Table 5 shows the results of our user-based evaluation. The average lexical precision achieved by the CSDOL system is 87.20% for concept learning, 91.44% for taxonomic relation learning, and 92.12% for association relation learning. The lexical precision of semantic relation learning was calculated based on the set of correctly extracted concepts by a system. The CSDOL system significantly outperforms the HK-means baseline system in domain concept learning ($t(9) = 303.1, p < 0.01$) and taxonomic relation learning ($t(9) = 51.6, p < 0.01$), respectively. Since the HK-means baseline system can only learn taxonomic relations, its lexical precision for association relation learning is not available. In this user-based semantic evaluation, the CSDOL system achieves an even higher average lexical precision for concept learning than that obtained when the Open Directory Project ontology is adopted as the standard. This is mainly because the user-contributed Open Directory Project ontology is not perfect. In some cases, the CSDOL system discovered correct concepts and semantic relations. However, these concepts and relations might be treated as mistakes when evaluated based on the Open Directory Project ontology. For example, the CSDOL system successfully learns the subconcept "self-service travel" under the concept "travel experience." However, such a taxonomic relation is not

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

Table 5 User-Based Evaluation Across Different Domains

	Lexical precision (CSDOL) (%)			Lexical precision (HK-means) (%)	
Domain	Concept	Taxonomic relation	Association relation	Concept	Taxonomic relation
Movie	87.43	91.43	92.57	62.29	65.71
Business	87.79	91.86	92.44	62.21	65.70
Video game	81.22	83.43	84.53	55.80	61.33
Software	88.54	93.23	93.75	63.54	67.19
Hardware	89.19	94.59	95.14	64.32	67.57
Medicine	87.86	93.06	93.64	63.01	66.47
Clothing	87.91	92.31	92.86	62.64	65.38
Travel	87.57	91.72	92.31	62.72	63.91
Cooking	87.01	90.96	91.53	62.15	63.28
AI	87.50	91.85	92.39	62.50	65.76
Average	87.20	91.44	92.12	62.12	65.23

captured by the existing Open Directory Project ontology. Appendix F of the online supplement shows the concept “travel experience” as learned by the CSDOL system and the HK-means baseline system, respectively.

6. Conclusions

Existing computational methods are weak in addressing the issues of social polysemy and social synonymy arising in ontology learning from folksonomies. This paper illustrates our cognitively motivated methodology, which exploits “contexts” and semantically rich concept vectors. To the best of our knowledge, this is the first research work that has successfully applied basic-level concept mining and a hybrid link- and content-based subsumption relation discovery method to learn context-sensitive domain ontologies from folksonomies. Our experimental results reveal that the proposed system significantly outperforms the best baseline ontology learning system by 13.83% in terms of taxonomic F-score. In addition, the domain ontologies generated by the proposed system are judged with high lexical precision by human experts. Our research work opens the door to automated learning of high-quality ontologies from folksonomies to support the retrieval of linked data on the Web and hence the interoperability among enterprises.

The limitations of the proposed methodology are that the basic-level concept-mining algorithm is not guaranteed to find the global optimum of category utility, and that contextual aspect mining requires a predefined number of latent topics as an input. Our future work will apply evolutionary computational methods to search for near-optimal category utility, number of latent topics, and other system parameters. Instead of using the current coarse-grained approach of representing contexts, future work will explore a taxonomy of contexts for the disambiguation of domain concepts and semantic relations. To improve the external validity of our research, we will examine the effectiveness of applying the automatically constructed domain ontologies to retrieve linked data on the Web.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/ijoc.2015.0644.

Acknowledgments

The work described in this paper was partly supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project CityU 145712] and grants from the Science Technology and Innovation Committee of Shenzhen Municipality—the Basic Research Program [Projects: JCYJ20130401145617281 and JCYJ20140419115614350]. Zhao’s work is partially supported by a grant from SAP Labs [Grant 9220059]. Cai’s work is supported by grants from NSFC [Grant 61300137], GNSFC [Grant S2013010013836], STPP of Guangdong Province [Grant 2013B010406004], and SCUT [Grant 2014ZZ0035].

References

Bizer C, Heath T, Berners-Lee T (2009) Linked data—The story so far. *Internat. J. Semantic Web Inform. Systems* 5(3):1–22.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

Borth D, Chen T, Ji R, Chang S-F (2013) Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. *Proc. 2013 ACM Multimedia Conf.* (ACM, Barcelona, Spain), 459–460.

Brice P, Jiang W, Wan G (2011) A cluster-based context-tree model for multivariate data streams with applications to anomaly detection. *INFORMS J. Comput.* 23(3):364–376.

Chun MM, Jiang Y (1998) Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psych.* 36(1):28–71.

Cimiano P, Hotho A, Staab S (2005) Learning concept hierarchies from text corpora using formal concept analysis. *J. Artificial Intelligence Res.* 24:305–339.

Daud A, Li J, Zhou L, Zhang L, Ding Y, Muhammad F (2010) Modeling ontology of folksonomy with latent semantics of tags. *Proc. IEEE/WIC/ACM Internat. Conf. Web Intelligence* (IEEE, Toronto, Canada), 516–523.

Downloaded from informs.org by [216.73.216.140] on 08 June 2026, at 13:16 . For personal use only, all rights reserved.

Gasevic D, Zouaq A, Torniai C, Jovanovic J, Hatala M (2011) An approach to folksonomy-based ontology maintenance for learning environments. *IEEE Trans. Learn. Tech.* 4(4):301–314.

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian relation of images. *IEEE Trans. Pattern Anal. Machine Intelligence* 6(6):721–741.

Gluck M, Corter J (1985) Information, uncertainty, and the utility of categories. *Proc. Seventh Annual Conf. Cognitive Sci. Soc., Berkeley, CA*, 283–287.

Gruber TR (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199–220.

Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. *Proc. 14th Internat. Conf. Comput. Linguistics, Nantes, France*, 539–545.

Heymann P, Garcia-Molina H (2006) Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford InfoLab, Stanford, CA.

Jupiter D, Şahutoğlu J, VanBuren V (2010) Treehugger: A new test for enrichment of gene ontology terms. *INFORMS J. Comput.* 22(2):210–221.

Lau RYK, Song D, Li Y, Cheung CH, Hao JX (2009) Towards a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE Trans. Knowledge Data Engrg.* 21(6):800–813.

Liu K, Fang B, Zhang W (2010) Ontology emergence from folksonomies. *Proc. 19th ACM Conf. Inform. Knowledge Management* (ACM, Ontario, Canada), 1109–1118.

Maedche A, Staab S (2004) Ontology learning. Staab S, Studer R, eds. *Handbook on Ontologies* (Springer, Berlin), 173–190.

Mika P (2007) Ontologies are us: A unified model of social networks and semantics. *J. Web Semantics* 5(1):5–15.

Missikoff M, Taglino F (2004) An ontology-based platform for semantic interoperability. Staab S, Studer R, eds. *Handbook on Ontologies* (Springer, Berlin), 617–634.

Munir MU, Javed MY, Khan SA (2012) A hierarchical *k*-means clustering based fingerprint quality classification. *Neurocomputing* 85:62–67.

Panetto H, Dassisti M, Tursi A (2012) ONTO-PDM: Product-driven ONTOlogy for product data management interoperability within manufacturing process environment. *Advanced Engrg. Informatics* 26(2):334–348.

Ponzetto SP, Strube M (2007) Deriving a large scale taxonomy from Wikipedia. *Proc. Twenty-Second National Conf. Artificial Intelligence* (AAAI Press/MIT Press, Cambridge, MA), 1440–1445.

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. *Cognitive Psych.* 8(3): 382–493.

Rosen-Zvi M, Chemudugunta C, Griffiths TL, Smyth P, Steyvers M (2010) Learning author-topic models from text corpora. *ACM Trans. Inform. Systems* 28(1):Article no. 4.

Roussinov D, Zhao JL (2003) Automatic discovery of similarity relationships through Web mining. *Decision Support Systems* 35(1): 149–166.

Salton G (1986) Recent trends in automatic information retrieval. *Proc. 9th Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 1–10.

Salton G, McGill MJ (1983) *Introduction to Modern Information Retrieval* (McGraw-Hill, New York).

Sanderson M, Croft B (1999) Deriving concept hierarchies from text. *Proc. 22nd Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 206–213.

Schmitz P (2006) Inducing ontology from Flickr tags. *Proc. 2006 Collaborative Web Tagging Workshop (WWW'06)* (ACM, Edinburgh, Scotland).

Strohmaier M, Helic D, Benz D, Körner C, Kern R (2012) Evaluation of folksonomy induction algorithms. *ACM Trans. Intelligent Systems Technol.* 3(4):Article no. 74.

Tanaka J, Taylor M (1991) Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psych.* 23(3): 457–482.

Tang J, Leung HF, Luo Q, Chen D, Gong J (2009) Towards ontology: Learning from folksonomies. *Proc. 21st Internat. Joint Conf. Artificial Intelligence, Pasadena, California*, 2089–2094.

Tao D, Li Y, Zhong N (2011) A personalized ontology model for Web information gathering. *IEEE Trans. Knowledge Data Engrg.* 23(4):496–511.

Tho QT, Hui SC, Fong ACM, Cao TH (2006) Automatic fuzzy ontology generation for semantic web. *IEEE Trans. Knowledge Data Engrg.* 18(6):842–856.

Trabelsi C, Jrad AB, Yahia SB (2010) Bridging folksonomies and domain ontologies: Getting out non-taxonomic relations. *The 10th IEEE Internat. Conf. Data Mining Workshops* (IEEE Computer Society, Sydney, Australia), 369–379.

Wasserman S, Faust K (1999) *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, UK).

Wei W, Barnaghi PM, Bargiela A (2010) Probabilistic topic models for learning terminological ontologies. *IEEE Trans. Knowledge Data Engrg.* 22(7):1028–1040.

Wong W, Liu W, Bennamoun M (2012) Ontology learning from text: A look back and into the future. *ACM Comput. Survey* 44(4):Article no. 20.

Yan X, Lau RYK, Song D, Li X, Ma J (2011) Towards a semantic granularity model for domain-specific information retrieval. *ACM Trans. Inform. Systems* 29(3):Article no. 15.

Zavitsanos I, Paliouras G, Vouros G (2011) Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Trans. Knowledge Data Engrg.* 23(11):1635–1648.

Zhou M, Bao S, Wu X, Yu Y (2007) An unsupervised model for exploring hierarchical semantics from social annotations. *6th Internat. Semantic Web Conf.*, Lecture Notes in Computer Science, Vol. 4825 (Springer, Busan, Korea), 680–693.

Zouaq A, Nkambou R (2009) Evaluating the generation of domain ontologies in the knowledge puzzle project. *IEEE Trans. Knowledge Data Engrg.* 21(11):1559–1572.