

Ontology Summarization Based on RDF Sentence Graph

Xiang Zhang
School of Computer Science
and Engineering
Southeast University
Nanjing 211189, P.R. China
xzhang@seu.edu.cn

Gong Cheng
School of Computer Science
and Engineering
Southeast University
Nanjing 211189, P.R. China
gcheng@seu.edu.cn

Yuzhong Qu
School of Computer Science
and Engineering
Southeast University
Nanjing 211189, P.R. China
yzqu@seu.edu.cn

ABSTRACT

Ontology summarization is very important to quick understanding and selection of ontologies. In this paper, we study extractive summarization of ontology. We propose a notion of RDF sentence as the basic unit of summarization. An RDF Sentence Graph is proposed to characterize the links between RDF sentences derived from a given ontology. The salience of each RDF sentence is assessed in terms of its “centrality” in the graph. We propose to summarize an ontology by extracting a set of salient RDF sentences according to a re-ranking strategy. We compare several measurements in assessing the salience of RDF sentences and give an overall evaluation of experiment results, which shows that our approach to ontology summarization is feasible.

Categories and Subject Descriptors

I.7.m [Computing Methodologies]: Document and Text Processing—*Miscellaneous*; I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods—*Representation languages*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Ontology Summarization, RDF Sentence, RDF Sentence Graph, Centrality, Re-ranking

1. INTRODUCTION

Ontologies have played a central role in the development and deployment of the semantic web. As defined by Gruber, an ontology is a formal explicit specification of a shared conceptualization [11]. From the viewpoint of knowledge engineers, ontologies embody formalized knowledge, which can be understood and reused by others.

Understanding of ontology is significant in the course of ontology development, semantic annotation and semantic data retrieval. When developing an ontology, a better choice for an engineer is to investigate and reuse existing ontologies rather than develop a new one from scratch. Ontology reuse leads to a network effect, which is called “ontology

convergence”. Ontology engineers need to understand available ontologies before reusing some of them. In the course of semantic annotation and semantic data retrieval, concepts defined in ontologies are instantiated to describe real-world objects and may be retrieved by users. The accuracy of retrieval is strongly affected by the annotators’ understanding of ontologies used.

The amount of available web ontologies continues increasing at a phenomenal rate. Swoogle [5] has indexed more than 10,000 ontologies and many more semantic web documents so far. Although the ontology search engines like Swoogle facilitate the finding and retrieval of existing ontologies, they do not ease the burden on users of understanding them for potential reuse. Just as search engines for web pages, a broad-topic query on ontology search engines often result in a ranked list of hundreds of ontologies.

Since text representation of expressive ontologies can not be friendly to human reading, most researches on facilitating the understanding of ontologies are focused on visualization technologies, which help users quickly understand the ontology by a graph presentation, or furthermore, allowing users to navigate the ontology in different level of details [8, 24, 28]. But there are also constraints with ontology visualization. A major constraint is the cost of communication. In a bandwidth-limited environment, text is preferable than graphics in the communication between agents. Thus, visualization technology is hardly incorporated into some information retrieval tasks.

The natural language processing, especially the automatic text summarization is widely used in understanding unstructured documents. Defined by Mani [18], the text summarization is “the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)”. Applications of text summarization are usual on the web, such as producing indicative summaries of web pages by search engines. The success of text summarization raise a question: is there a feasible approach to ontology summarization?

It motivates us to find a new method to facilitate the understanding and selection of ontologies. Firstly, the method should provide the user with a concise and indicative representation of ontology for a quick understanding. Secondly, the summarization should be coherent and has a extensive coverage on multiple subjects of the ontology. At last, transferring such representation of ontology on the Web should be low-cost to be incorporated into ontology retrieval tasks.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

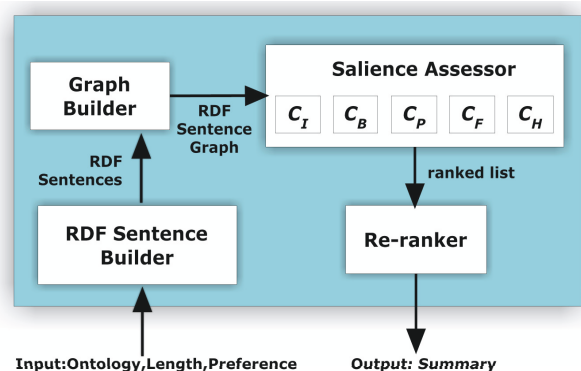


Figure 1: Architecture of ontology summarization

In this paper, we give an answer to this question. In Section 2, we give an overview of our approach to ontology summarization. A formal definition on “RDF sentence” is proposed in Section 3, which has a similar structure with sentences in natural language. We characterize the navigation behavior of a surfer on an “RDF Sentence Graph”. In Section 4, we elaborate several possible measurements to assess the salience of RDF sentences based on their centralities in the RDF Sentence Graph. In Section 5, we present a re-ranking algorithm to extract salient RDF sentences into ontology summaries. In Section 6, we give an evaluation on our approach based on “ground truth” summaries produced by human experts. Related works are discussed in Section 7, and we make a conclusion in the last section.

2. OVERVIEW

Ontology summarization is the process of distilling knowledge from ontology to produce an abridged version for a particular user (or users) and task (or tasks).

It is argued that RDF statements can be the basic unit in the process of distilling. However, extracted summaries in the form of RDF statements will encounter a blank node problem: it is possible that RDF statements sharing common blank nodes are separated, which could lead to a summary containing some meaningless RDF statements. RDF sentences provide integrated information, which is a combination of a set of RDF statements connected by blank nodes. Thus, they “encapsulate” the blank nodes in local structures. Using the notion of RDF sentence as a basic analytic unit, the shortcoming of RDF statements will be overcome. Further, an ontology (including RDFS and OWL ontologies) can be viewed as linked RDF sentences, which leads to an RDF Sentence Graph. A formal definition of RDF sentence and RDF Sentence Graph will be given in the next section.

Figure 1 exhibits the architecture of our ontology summarization, which is composed of four major components:

RDF SENTENCE BUILDER: This component accepts user-provided ontology. Usually, a user may also provides the expected length of the final summary, and he is allowed to further customize his navigational preference, which will be used to determine the weight of links between RDF sentences. The ontology is mapped to a set of RDF sentences.

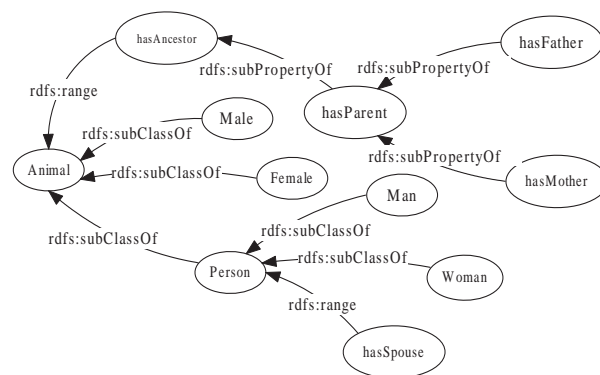


Figure 2: Graph summary of the Animal Ontology

GRAPH BUILDER: Based on the set of RDF sentences and user’s preference, an RDF Sentence Graph is build, which characterizes the links between RDF sentences in the ontology. We will give a formal definition of RDF Sentence Graph in Section 3.

SALIENCE ASSESSOR: This component assesses the salience of RDF sentences based on link analysis of RDF Sentence Graph. In our approach, the salience of an RDF sentence is determined by the centrality of corresponding vertex in RDF Sentence Graph. Since the RDF Sentence Graph carries the user’s preference, the salience of RDF sentence indicates its importance in the ontology from the user’s view. Five different centrality measurements are elaborated in Section 4. The Salience Assessor finally gives a ranking to RDF sentences according to their salience.

RE-RANKER: This component produces the final summary of the ontology. It does not simply extract the user-specified number of most salient RDF sentences. The coherence of the summary and its coverage on the original ontology are also considered. Salient RDF sentences are re-ranked before extracted into the summary. A re-ranking algorithm will be given in Section 5.

We offer an online service of our ontology summarization. The portal of our service can be found at: <http://iws.seu.edu.cn/services/falcon-f/ontosum/>. A user can upload his ontology and specify the expected length of the summary. Result summaries are presented in text and graph. Text summaries are sequences of sentences and namespaces are omitted; graph summaries are visualizations of text summaries, which are SVG figures. A graph summary of the Animal Ontology (<http://www.atl.lmco.com/projects/ontology/ontologies/animals/animalsA.owl>) is shown in Figure 2. This ontology is also a test case in the evaluation section.

3. LINKED RDF SENTENCES

3.1 RDF Sentence

An RDF graph $G(T)$ can be mapped into a set of RDF statements T , composed of URI references, literals and blank nodes, making descriptions about resources. According to the RDF semantics, blank node is a kind of existentially

quantified resources whose meaning is in the scope of the graph it occurs. RDF statements sharing a common blank node form a structure providing a joint context of the blank nodes. If such RDF statements are separated into different graphs, the context is broken. The structure is important to certain applications, which will reference or extract a sub-graph of an RDF graph and meanwhile require the extraction to retain meaningful. However, RDF semantics does not provide any intrinsic mechanism to identify this kind of structure. We define it as an *RDF Sentence*:

B-CONNECTED: We say two RDF statements are b-connected if they share common blank nodes. Besides, the b-connected relation is transitive, i.e., two RDF statements are said to be b-connected if they are both b-connected to another RDF statement. To guarantee that the b-connected RDF statement are always grouped together, we introduce the notion of RDF sentence, which corresponds to a maximum set of b-connected RDF statements. The formal definition is given as follows.

Definition 1. (RDF SENTENCE): For an RDF graph $G(T)$, an RDF sentence s is a set of RDF statements, which satisfies the following conditions:

1. $s \subseteq T$;
2. $\forall i, j \in s, i, j$ are b-connected;
3. $\forall i \in s, j \notin s, i, j$ are not b-connected.

We call an RDF statement whose subject is not a blank node as a main RDF statement. Generally, an RDF sentence is formed by a main RDF statement and all the other RDF statements b-connected to it. We name this kind of RDF sentence as a “generic” RDF sentence. Sometimes however, an RDF sentence may have zero or more than one main statements, we distinguish this kind of RDF sentence as “special” RDF sentence.

An OWL DL ontology can be mapped to a corresponding RDF graph [20], and further to a set of RDF sentences. In OWL abstract syntax, facts or axioms of atomic class (property) are usually mapped to “generic” RDF sentences. Some OWL DL Axioms are mapped to “special” RDF sentences, such as axioms specifying the equivalence between class restrictions. While such axioms are important for the tasks of reasoning, they are less important to certain tasks such as ontology summarization. Besides, they will break the simplicity of our model, which will be described later. Therefore, when we refer to the notion of RDF sentence hereinafter, we actually refer to “generic” RDF sentences. “Special” RDF sentences are ignored in ontology summarization. We also assume that users are only interested with RDF sentences at concept level, which are sentences about classes and properties. Other sentences, such as ontology header and RDF sentences about instances are also ignored. To facilitate later discussion, we define the Subject / Predicate / Object of an RDF sentence as follows.

Definition 2. (SUBJECT / PREDICATE / OBJECT OF AN RDF SENTENCE): The subject and predicate of a generic RDF sentence are the subject and predicate of its main RDF statement; the object of a RDF sentence is the set of terms occurring in the RDF sentence, except the subject and predicate of the main RDF statement.

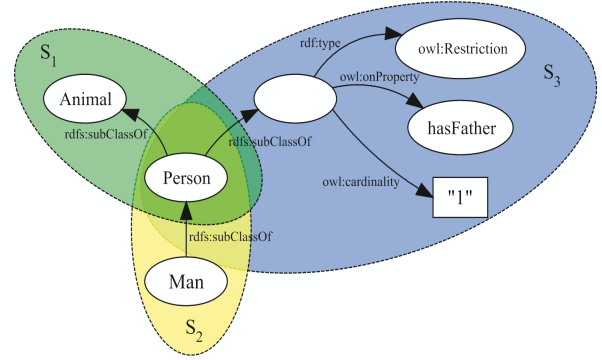


Figure 3: A graph representation of three RDF sentences derived from the Animal Ontology

Here, “term” refers to URI reference which is defined by users, not belonging to the build-in vocabulary of ontology language.

An example is shown in Figure 3, which is a subgraph of the RDF graph derived from the Animal Ontology. The graph can be divided into three RDF sentences: S_1 , S_2 and S_3 .

3.2 Links between RDF Sentences

Natural language sentences are sequential in text, while RDF sentences can be structured as a graph. It is natural to define the links between RDF sentences based on common terms they shared. The links can be classified into two classes: sequential or coordinate, depending on the position of common terms.

SEQUENTIAL LINK: There is a sequential link from one RDF sentence to another if the predicate or an element in the object of the former is the same term with the subject of the latter. We name this type of link as sequential since it represents the relation similar to the sequential relation between natural language sentences.

COORDINATE LINK: There is a coordinate link from one RDF sentence to another if the subject of both RDF sentences are the same. We name this type of link as coordinate since it represents the relation similar to the coordinate relation between natural language sentences, which often appears as a compound sentence.

3.3 RDF Sentence Graph

Considering a scenario that a user is navigating inside an ontology, when he reads an RDF sentence, he wants to further look up some RDF sentences concerning the object of the current one. In this case, he will follow a sequential link; or he may want to read more RDF sentences talking about the subject of the current one, and then follows a coordinate link. The user may have a preference on which type of link to follow. We characterize the preference by a parameter p , which is a value between 0 and 1 representing the probability of following sequential links, and thus $1 - p$ of following coordinate links. We define an RDF Sentence Graph, which is a weighted and directed graph, characterizing the links between RDF sentences from the viewpoint of a user.

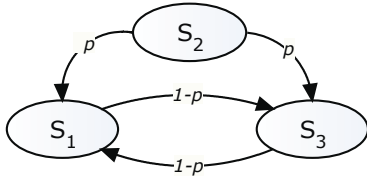


Figure 4: An RDF Sentence Graph

Definition 3. (RDF SENTENCE GRAPH): An RDF Sentence Graph $G = (V, E, W, p)$ is a weighted and directed graph, where each vertex represents an RDF sentence. Given $i, j \in V$ and $i \neq j$, $(i, j) \in E$ iff there exists at least one sequential link or coordinate link from i to j , where V and E stand for the set of vertices and edges. W is the set of edge weights, and each weight is calculated as:

$$w(i, j) = p * seq(i, j) + (1 - p) * cor(i, j),$$

where $seq(i, j)$ and $cor(i, j)$ are 0 or 1, representing the existence of sequential or coordinate link from i to j respectively, and p is the navigational preference of a surfer.

From the definition of RDF Sentence Graph, we can see it carries the preference of a user, which can be a person or software agent. RDF Sentence Graph is customizable, and different users could have different RDF Sentence Graphs for a given ontology. A default preference can be defined based on a generally accepted navigational preference, which will be discussed in the evaluation section.

The RDF Sentence Graph built from RDF sentences in Figure 3 is shown in Figure 4.

4. SALIENCE OF RDF SENTENCE

A naive method to assess the salience of RDF sentences can be rooted from the idea of “centroid” in text summarization [21]: the salience of RDF sentences can be determined by the salience of terms referred by them. Treating the name of each term as a sequence of words and the ontology as a bag of terms, the salience of terms can be assessed by comparing to the “centroid” of the ontology, which is a pseudo-document consisting of words whose frequencies are above a predefined threshold.

Although the “centroid”-based salience is easy to assess, the method ignores the graph nature of ontologies. In our approach, the salience of an RDF sentence is assessed in terms of its “centrality” in the RDF Sentence Graph. The notion of “centrality” defined on the vertices of a graph is a measurement originated from the analysis of social networks. They are designed to rank the actors according to their positions in the network and interpreted as the salience of actors embedded in a social structure.

In this section, we give a survey on different notions of centrality and their measurements, which can be classified into three categories: Degree Centrality, Shortest-Path-Based Centrality and Eigenvector Centrality. Before we make a conclusion that which measurement is suitable to assess the salience of RDF sentences, five different measurements are selected from the three categories. They are well-studied and can be applied to weighted and directed graphs such as the RDF Sentence Graphs. We distinguish the intuitive meaning of the “salience” assessed by each of them in the

following subsections, and then make a comparison among them in the evaluation section.

It is notable that since the salience of an RDF sentence is assessed based on the RDF Sentence Graph, it has a bias to the navigational preference of the user.

Our work is related to a graph-based text summarization described in [7] and a semantic network analysis on ontologies discussed in [13]. The former measures the centrality of sentences on a graph of lexical cohesion, and the latter measures the centrality of inter-connected terms in ontologies.

4.1 Degree Centrality

Degree centrality is a simple measurement of vertices’ salience. In an undirected graph, the degree centrality of a vertex is measured by the number of connections it has. For a directed graph, the degree centrality of a node is measured by the number of its incoming links, which is called in-degree centrality; or by the number of its outgoing links, which is called out-degree centrality.

Links between vertices can be seen as conferral of authority. Vertices with high degree centrality are intuitively salient in the graph since they receive many conferral of authority from others. The conferral can be discrete for unweighted graph; or continuous for weighted graphs.

Given a weighted directed graph $G = (V, E, W)$, where V is the set of vertices, E is the set of edges, W is the set of edge weights. C_I and C_O of each vertex are computed as following:

$$C_I(i) = \sum_{(j,i) \in E} w(j,i),$$

$$C_O(i) = \sum_{(i,j) \in E} w(i,j).$$

In this category, we select weighted in-degree (C_I) centrality to assess the salience of RDF sentence in an RDF Sentence Graph. A salient RDF sentence with a high C_I indicates that many other RDF sentences have links to it.

4.2 Shortest-Path-based Centrality

In social network analysis, many notions of centrality are based on shortest paths linking pairs of actors. For example, the salience of an actor can be measured by the maximum or total distance from it to others, which are called Graph Centrality [12] denoted by C_G and Closeness Centrality [22] denoted by C_C respectively; or the ratio of shortest paths across the actor in the network, which is called Betweenness Centrality [10] denoted by C_B .

Shortest-path-based centrality characterizes the position of a vertex in a graph, which brings more information than degree centrality. We select the betweenness centrality (C_B) in this category since it provides more topological information than others, which is essential in the analysis of social networks. The original notion of betweenness centrality is defined on undirected and unweighted graph. The notion is generalized to directed cases in [29]. An algorithm is introduced in [1] to efficiently compute the betweenness centrality in directed and weighted graphs.

Salient RDF sentences with high betweenness centrality can be seen as “bridges” between clusters of RDF sentences: the “bridge” RDF sentences have direct links to lots of sentences in clusters, while few links lie between sentences belonging to different clusters.

4.3 Eigenvector Centrality

Graphs can be encoded in adjacency matrices. The entries in the matrix are either 1 if a connection exists between two vertices, or 0 if not. The matrix is symmetric if the connection is undirected or asymmetric otherwise.

Two well-known measurements of eigenvector centrality on the Web is PageRank [19] and HITS [16]. PageRank is used by the Google search engine for ranking web pages. The authority of a page is computed recursively as a function of the authorities of the pages that link to it. HITS computes two values related to topological properties of the Web pages, the “authority” and the “hubness”. The authorities indicates the page relevance as information source, which are parallel to the main eigenvector of the bibliographic coupling matrix; while the hubness refers to the quality of a page as a link to authoritative resources, which are parallel to the main eigenvector of the co-citation matrix [15].

An unified probabilistic framework has been introduced in [3] to interpret PageRank, HITS and other web page scoring algorithms using a random walk model. While the original PageRank and HITS algorithms are dealing with directed but unweighted graph such as the web graph, the probabilistic framework extends both algorithms, by which link weights affect the surfers’ actions when they walk on the graph. We use three weighted variations of PageRank and HITS to define the eigenvector centrality of RDF sentences.

Weighted PageRank and Focused Weighted PageRank

Imagining a scene of a user’s random walk on an RDF Sentence Graph, the user can make a choice from two atomic actions when he is staying at some page: forward a link from current sentence or jump to another. A Markov chain can be derived from the random walk, characterizing the probability of transition from one RDF sentence to another at a certain step. If the user forwards or jumps to other sentences randomly, the stationary distribution of the Markov chain results to be the original PageRank values of RDF sentences when the steps of the user tend to infinite.

In another case, the user doesn’t forward a random link. Instead, he forwards each link with same or different probability according to the link weights. We call this model as “Weighted PageRank”. We denote the centrality of RDF sentences measured by Weighted PageRank as C_P . Weighted PageRank is similar to “Focused PageRank” described in [3]. While the Weighted PageRank is generic, the Focused PageRank is topic-sensitive. Furthermore, if the user also jumps to other sentences non-randomly, and the probability of jumping is determined by the similarity of the sentence to a certain topic, we call this model “Focused Weighted PageRank”, and denote corresponding centrality as C_F . The C_P and C_F of each RDF sentence can be computed as following, where d is the damping factor; $s(i)$ is the similarity of RDF sentence i to a certain topic.

$$C_P(i) = \frac{1-d}{N} + d * \sum_{(j,i) \in E} \frac{w(j,i) * C_P(j)}{\sum_{(j,k) \in E} w(j,k)},$$

$$C_F(i) = \frac{1-d}{N} * \frac{s(i)}{\sum_{p \in V} s(p)} + d * \sum_{(j,i) \in E} \frac{w(j,i) * C_F(j)}{\sum_{(j,k) \in E} w(j,k)}.$$

C_P and C_F are both selected to assess the salience of RDF sentences in an RDF Sentence Graph. It is notable that in the Focused Weighted PageRank, we define the $s(i)$ as the

similarity between the Virtual Documents (VDoc in short) of an RDF sentence and the ontology. The notion of VDoc is defined in [27], which is a collection of weighted words containing the local descriptions and neighboring linguistic information of a term to reflect the intended meaning of the term. In an ontology, local descriptions include local names, labels, comments and other annotations of terms. Here, the VDoc of an RDF sentence is the combination of VDocs of the terms it refers; and the VDoc of an ontology refers to the combination of VDocs of its vocabulary. A similar approach was introduced in [30] for assessing the salience of terms. The similarity between an RDF sentence and the ontology can be seen as a “linguistic centrality”, and Focused Weighted PageRank gives us a powerful tool to combine both structure and linguistic information for assessing the salience of RDF sentences.

Salient RDF sentence with a high C_P indicates that many other salient RDF sentences have links to it. A high C_F further indicates that the RDF sentence is also salient in the linguistic level.

Weighted HITS

In previous centrality measurements, we assess the salience of RDF sentence regardless of the type of its subject. In other words, RDF sentences about classes or properties are treated equally. However, user may focus on their different topological properties. Generally, a user may think that an RDF sentence about a property is salient if it links to many salient RDF sentences about classes; and similarly, an RDF sentence about a class is salient if it is linked from many salient RDF sentences about properties. It motivates us to analyze the hubness and authority of an RDF sentence. The salience of an RDF sentence is assessed by its hubness if its subject is a property; or the salience is assessed by its authority if its subject is a class.

Here we leave out the formulation of computing authority and hubness of vertices in a weighted graph, which are also introduced in [17] by using a probability framework. It is proved that the authority of a vertex in a graph is proportional to the sum of its weighted in-degree, and the hubness is proportional to the sum of its weighted out-degree, which suggests a very simple algorithm for calculating the centrality of RDF sentences. We denote this centrality measurement as C_H .

5. RE-RANKING

Recall the requirement of a good ontology summarization given in the introduction section. Ontology summaries should be concise and indicative. This requirement is satisfied by extracting the salient part of the ontology. Ontology summaries should also be expressed in a coherent way and has an extensive coverage over the multiple parts of the ontology. An extensive “coverage” means the summary should be an extraction reflecting diverse subjects of the original ontology and avoid extracting identical subjects, which can be seen as redundancy. “Coherent” here means extracted RDF sentences in the summary should not be irrelative, and their relation should be explicitly exhibited in the summary. However, this requirement cannot be satisfied by simply extracting the most salient part of the ontology.

The problem of information redundancy has been discussed in text summarization. Extracted text summaries are often redundant if many salient sentences talk about

similar things in the original text. A well-known solution to reduce information redundancy in text summaries has been introduced in [2], which is called Maximal Marginal Relevance (MMR). It is a query-related re-ranking algorithm, based on the idea that given a ranked list of sentences, they will be extracted into the summary according to a combined criterion of query relevance and novelty of information comparing to the already extracted sentences in the summary. The novelty of information can be inversely measured by a linguistic similarity metric, such as a cosine similarity. Another re-ranking algorithm in text summarization has been stated in [21], which consider the information subsumption among sentences. Motivated by these works, we design a new re-ranking algorithm to reduce redundancy in ontology summaries, and meanwhile enhance the coherence of the summaries.

Let R represents a ranked list of RDF sentences derived from an ontology, which is produced by the Saliency Assessor; S represents a set of already extracted RDF sentences in the summary; i is the index of the next RDF sentence to be extracted. i should satisfy that:

$$i = \arg \max_{i \in R \setminus S} [C(i) + \lambda * reward(i, S) - (1 - \lambda) * penalty(i, S)],$$

where $C(i)$ is the centrality of i ; $reward(i, S)$ is a metric measuring the potential contribution of i to the total coherence of the summary if i is extracted and added to S in the next step; $penalty(i, S)$ measures the guilt of i to the increase of redundancy if i is extracted and added to S . λ is a parameter controlling the effect of reward and penalty. A related discussion of the chosen re-ranking parameters in MMR has been introduced in [2], and we simply choose $\lambda = 0.5$.

Intuitively, if many sequential links lie between RDF sentences in a summary, we consider the summary is coherent because it presents a sequence of knowledge which makes the summary comprehensible for human reading; if many coordinate links lie between RDF sentences in the summary, we consider the summary is highly redundant, since a major part of the summary is talking about a same subject. Based on this intuition, we define the two metrics as following, where $seqs$ and $cors$ are the number of sequential or coordinate links in S respectively.

$$\begin{aligned} reward(i, S) &= 1 - \frac{seqs}{seq_{S \cup \{i\}}}, \\ penalty(i, S) &= 1 - \frac{cors}{cor_{S \cup \{i\}}}. \end{aligned}$$

6. EVALUATION

Although there is no other published work on ontology summarization to be compared with ours, evaluation of text summaries has been discussed for years. Generally, summarization technologies are evaluated by measuring the agreement of their extracted summaries to human-generated summaries, which are called “ground truths”. Many evaluation metrics had been proposed to measure the performance of a summarization approach. According to [6], these metrics can be classified into three categories: recall-based, sentence-rank-based and content-based.

In this section, we present the evaluation of our ontology summarization, based on ground truth summaries produced by human experts. We draw a lot of experiences from eval-

Table 1: Statistical data of the test cases

	RDF statements	RDF sentences	Classes	Properties
Animal	129	103	9	15
Beer	169	169	51	12
CV	416	416	16	72

uation metrics of text summaries. Three different metrics are used in different evaluation tasks.

6.1 Case Study

We have selected three small ontologies as the test cases. They are the Animal ontology (<http://www.atl.lmco.com/projects/ontology/ontologies/animals/animalsA.owl>), the Beer ontology (<http://www.purl.org/net/ontology/beer.owl>), and the CV ontology (<http://captsolo.net/semweb/resume/cv.rdfs>). The Animal ontology describes a conceptual framework of person and relations between persons, such as parent relation and spouse relation. The Beer ontology models types of beer, brewers/brands and ingredients of beer. The CV ontology models basic elements of resume, such as the work history, education, skills and target. Both the Animal ontology and the Beer ontology are in OWL, and the CV ontology is in RDFS. The three ontologies are selected as the test case since they are rather small, which can be reviewed by human to produce ground truths.

Some statistical data of the test cases are shown in Table 1, including the total number of RDF statements and RDF sentences, and the number of classes and properties.

6.2 Evaluation Metrics

We invited five judges, who are all experts on Semantic Web, to take a peer review on each test cases. Each judge was required to give a score ranging from 1 to 5 to express his judgment on the salience of every RDF sentence, where score “1” represents “least salient” and “5” represents “most salient”. Then, each judge was asked to extracted 10 RDF sentences for each test case, which they believed are qualified to form a summary based on their own criterion. The ranking of RDF sentences produced by the scores and the human-generated summaries are the ground truths for the evaluation.

Kendall’s tau Statistic [23] (tau in short) is often used to measure the agreement between rankings of sentences produced by different summarization systems. This metric is a sentence-rank-based metric. In fact, the result of it does not reflect the practical performance of a summarization, but is often used to relatively compare different summarization systems. We use tau mainly to evaluate the ranking of RDF sentences produced by different centrality measurements.

Recall-based metrics are intuitive and simple, but it has been well-discussed that they are also limited in exhibiting the real performance of summarization systems mainly because of the “disagreement due to synonymy” [6]. Content-based metrics compute the similarity between two summaries at a finer grained level than recall. Formally, the performance of summarization systems can be measured using following formula, where M is a similarity metric, S is the summary produced by machine, $J_1 \dots J_n$ is human summaries:

$$sim(M, S, \{J_1 \dots J_n\}) = \frac{M(S, J_1) + M(S, J_2) + \dots + M(S, J_n)}{n}.$$

Table 2: Agreement between ground truths measured by tau

	J_1	J_2	J_3	J_4	J_5	avg.
Animal	0.360	0.225	0.302	0.132	0.121	0.228
Beer	0.483	0.549	0.528	0.540	0.484	0.517
CV	0.347	0.297	0.274	0.329	0.398	0.329

Table 3: Agreement between ground truths measured by vocabulary overlap

	J_1	J_2	J_3	J_4	J_5	avg.
Animal	0.813	0.750	0.750	0.611	0.700	0.725
Beer	0.614	0.686	0.750	0.673	0.708	0.686
CV	0.500	0.422	0.634	0.634	0.646	0.567

Vocabulary Overlap is a widely used content-based metrics in text summarization, where the similarity between summaries is measured by the overlap of vocabularies using cosine similarity or n-gram. For a better evaluation of ontology summarization, we also define a novel vocabulary overlap metric, where the similarity measure is:

$$M(S, J) = \frac{|T_{complete}| + 0.5 * |T_{partial}| + 0.25 * |T_{sibling}|}{|T|}$$

In above formulism, T is the set of terms contained in J , $T_{complete} \subseteq T$, which is the set of terms having identical or equivalent terms in S ; $T_{parital} \subseteq T - T_{complete}$, which is the set of terms having super or sub-terms in S ; $T_{sibling} \subseteq T - T_{complete} - T_{parital}$, which is the set of terms having direct sibling terms in S .

We use vocabulary overlap to measure the quality of ontology summaries produced by our approach and the performance of re-ranking algorithm.

6.3 Analysis of Ground Truths

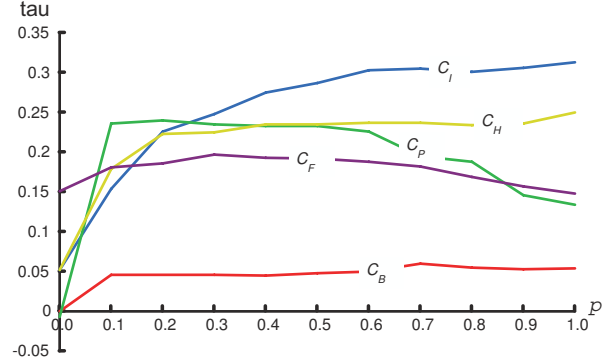
Firstly, we give an evaluation on the agreement between ground truths. By assigning a score to indicate the salience of every RDF sentence, each judge actually ranked RDF sentences into five classes from least to most salient. Table 2 shows the agreement among judges on the salience of RDF sentences. Each entry is an average correlation between a judge to any other, which is measured by Kendall's tau Statistic. A total average correlation for each test case is shown in the rightmost column.

We can see that the average ranking correlation of Animal and CV ontology are both less than 0.5. It shows that disagreements are frequent among judges. However, there is a great vocabulary overlap between human summaries, which is shown in Table 3. It is partially caused by the fact that each test case has a small vocabulary. Another important reason is that: although judges have disagreement on the salience of some RDF sentences, they usually agree on the major topic of the ontology. Therefore, the summary they extracted often contains common terms.

We observed that nearly 10 percents of RDF sentences in human summaries are not extracted from the top 10 salient RDF sentences according to the scores given by judges. The explanations of judges are the same: to avoid redundancy in the summary. That is also the same motivation of our re-ranking algorithm.

Table 4: Agreement between various centrality measurements measured by tau

	C_I	C_B	C_P	C_H	C_F
C_I	1.000				
C_B	0.393	1.000			
C_P	0.681	0.281	1.000		
C_H	0.669	0.462	0.480	1.000	
C_F	0.600	0.274	0.725	0.347	1.000

**Figure 5: Performance of various centrality measurements with different p**

6.4 Evaluation of Centrality Measurements

In Section 4, we have selected five different centrality measurements. We now give a comparison between them, and then evaluate their performance on assessing the salience of RDF sentences.

Different centrality measurements will result in different rankings of RDF sentences. In Table 4, we exhibit the agreement between rankings produced by each measurement on each test case, where the navigational preference p is set to 0.5.

It is clear that weighted in-degree (C_I) produces a similar ranking with weighed PageRank (C_P), weighted HITS (C_H) and focused weighted PageRank (C_F), which is natural because their intuitive interpretations are similar. Betweenness centrality (C_B) produces a ranking that is unlike to others. As stated previously, we observe that salient RDF sentences measured by C_B do play a “bridge” role in the ontology, such as RDF sentences that are domain and range specification of a property, which often link to RDF sentences about classes in isolated class hierarchy.

Figure 5 shows the agreement of ranking produced by each measurement with the ranking produced by human judge under different navigation preference.

It is a little surprising that in this test, C_I seems to be the best choice to assess the salience of RDF sentences, although it is the simplest. Its correlation to human judgment is higher than others, and even comparable to correlation among human judgments. The next choice will be C_H or C_P . But the curve of C_P keeps descending when p increases, while curves of other measurements keeps ascend or stable. The reason is that: many RDF sentences about properties act as hubs, linking to RDF sentences about classes. In the model of C_P , they confer authority to others but receive none. The increasing of p has the effect that they will con-

fer more authority while still receive none. It makes them "least salient" in the ontology, which is not consistent with human judgement. This also explains why the curve of C_H is stable. C_B has the worst performance comparing to others. We can conclude that the "bridge" RDF sentences are not salient according to the human's understanding.

When $p = 0$, all the measurements have a poor correlation with human judgment, in which C_F is the best among them. It can be interpreted as following: $p = 0$ means the sequential links are ignored. It makes the RDF Sentence Graph separated into a lot of isolated subgraphs, which can be seen as definitions of terms. The great information loss results in the inefficiency of all the network analysis methods. In this case, the ontology summarization becomes a random extraction, which makes the correlation close to zero. But C_F is partially determined by the linguistic information of RDF sentences, which will not be affected by users' navigational preference. Another interesting result is that: when $p = 1$, all the measurements are still effective, which indicates that ignoring coordinate links does not necessarily separate the ontologies into many isolated subgraphs.

From this test, we can make the conclusion that, both degree centrality and eigenvector centrality are suitable to assess the salience of RDF sentence. Among them, weighed in-degree is the best choice since its simplicity. And from the curves, we observed that, the generally accepted navigational preference p seems to be a relatively large number. In latter evaluation, we set it to 0.8.

6.5 Evaluation of Ontology Summaries

To evaluate the quality of our machine-generated summaries, each test case is distilled into five summaries using five centrality measurements. Each summary contains 10 extracted RDF sentences, and is compared to summaries of ground truths by measuring the average vocabulary overlap between them. The evaluation result is shown in Table 5. It is obvious that a relatively high vocabulary overlap exists between summaries produced by C_I , C_P and C_H and human summaries, which are comparable to the overlaps among human summaries as shown in Table 3.

It is also interesting that after re-ranking, the quality of summaries produced by C_I , C_P and C_H are close, and summaries produced by C_P have the best quality. We observe that although the ranking of RDF sentences produced by C_P is not the best comparing to ground truths, it usually agree with them on "most salient" RDF sentences. The summaries produced by our approach are actually determined by some "most salient" RDF sentences together with the re-ranking algorithm, which improves the performance of C_P in this test. It also indicates that the re-ranker has a great impact on our ontology summarization approach.

We also compare the quality of summaries produced by our approach with or without applying the re-ranking algorithm, which is exhibited in Table 6. We can see a promotion of agreement after re-ranking. Due to the limitation of space, we only show the evaluation of re-ranking based on C_P . In fact most centrality measurements produce better summaries by applying the re-ranking algorithm.

7. RELATED WORK

There is no comparable published work on ontology summarization. But summarizing ontology has been used in some reasoning tasks. Fokouel et al. [9] proposed an ap-

Table 5: Evaluation of summarization quality measured by vocabulary overlap

	Animal	Beer	CV
C_I	0.626	0.651	0.403
C_B	0.415	0.387	0.293
C_P	0.650	0.691	0.500
C_H	0.598	0.601	0.492
C_F	0.573	0.621	0.310

Table 6: Evaluation of re-ranking measured by vocabulary overlap (C_P)

	Animal	Beer	CV
Without re-ranking	0.496	0.657	0.251
With re-ranking	0.650	0.691	0.500

proach to summarize Abox in secondary storage by reducing redundancy to make reasoning scalable for very large Aboxes. It is an alternative approach with KAON2 [14], which reduces an *SHIQ*(\mathcal{D}) ontology to a disjunctive datalog program and makes it naturally applicable to Aboxes stored in deductive databases.

The notion of RDF sentence is originated from [26], where triples with blank nodes are divided into groups according to an equivalence closure for constructing knowledge base from RDF graph. A similar notion is stated in [25], which is called self Minimum Self-contained Graph providing a unit of RDF graph for signing; meanwhile, a notion of RDF molecule is proposed in [4], which is a trackable unit providing provenance information. In [30], we give an early description of the RDF sentence, and use it as an indication of the dependency between domain entities within ontology.

8. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel approach to automatic ontology summarization based on RDF Sentence Graph. Summaries are customizable: users can specify the length of summaries and their navigational preferences. We compared five different centrality measurements in assessing the salience of RDF sentence and defined a reward-penalty re-ranking algorithm to make the summaries comprehensive.

The evaluation showed that weighted in-degree centrality measures and several eigenvector centralities all have good performance in producing qualified summaries after re-ranking. Shown by the experiments, our approach of ontology summarization is feasible and promising.

In future work, we are going to explore our approach in real-world applications, such as ontology indexing and retrieval. We will exploit the approach of indexing the summary of ontologies rather than indexing the original ontologies. Besides, our current work is topic-independent, which provides generic information of ontology and aims at a broad readership. It is also interesting to consider a topic-dependent ontology summarization.

Acknowledgements

The work is supported by the 973 Program of China under Grant 2003CB317004. The third author of this paper is also supported by NCET (New Century Excellent Talents

in University) Program under Grant NCET-04-0472. We would like to thank Hongda Li and Dao Yin for their work on related experiments. We are also grateful to Wei Hu for his valuable suggestions.

9. REFERENCES

- [1] Brandes, U. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2) (2001), 163–177
- [2] Carbonell, J., and Goldstein, J. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1998), 335–336
- [3] Diligenti, M., Gori, M., Maggini, M. A Unified Probabilistic Framework for Web Page Scoring Systems. *IEEE Transaction on Knowledge Data Engineering*, 16(1) (2004), 4–16
- [4] Ding, L., Finin, T., Peng, Y., Joshi, A., da Silva, P., McGuinness, D. Tracking RDF Graph Provenance using RDF Molecules. In *Proc. of the 4th International Semantic Web Conference (Poster)*, (2005)
- [5] Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P. Finding and Ranking Knowledge on the Semantic Web. In *Proc. of the 4th International Semantic Web Conference*, (2005), 156–170
- [6] Donaway, R.L., Drummey, K.W., Mather, L.A. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proc. of ANLP/NAACL Workshop on Automatic Summarization*, (2000), 69–78
- [7] Erkan, G., and Radev, D.R. LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization. *Journal of Artificial Intelligence Research*, 22 (2004), 457–479
- [8] Fluit, C., Sabou, M., van Harmelen, F. Supporting User Tasks through Visualisation of Light-weight Ontologies. *Handbook on Ontologies in Information Systems*, (2003), 415–434
- [9] Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K. The Summary Abox: Cutting Ontologies Down to Size. In *Proc. of the 5th International Semantic Web Conference*, (2006), 343–356
- [10] Freeman, L.C. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40 (1977), 35–41
- [11] Gruber, T.R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2) (1993), 199–220
- [12] Hage, P., and Harary, F. Eccentricity and Centrality in Networks. *Social Networks*, 17 (1995), 57–63
- [13] Hoser, B., Hotho, A., Jaschke, R., Schmitz, C., Stumme, G. Semantic Network Analysis of Ontologies. In *Proc. of the 3rd European Semantic Web Conference*, (2006), 514–529
- [14] Hustadt, U., Motik, B., Sattler, U. Reducing SHIQ Description Logic to Disjunctive Datalog Programs. In *Proc. of the 9th International Conference on Knowledge Representation and Reasoning*, (2004), 152–162
- [15] Kessler, M.M. Bibliographic Coupling between Scientific Papers. *American Documentation*, 14 (1963), 10–25
- [16] Kleinberg, J. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms*, (1998), 668–677
- [17] Lempel, R. and Moran, S. The Stochastic Approach for Link-structure Analysis (SALSA) and the TKC Effect. In *Proc. of the 9th International World Wide Web Conference*, (2000), 387–401
- [18] Mani, I. Automatic Summarization. John Benjamins Publishing Company, (2001)
- [19] Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University, (1998)
- [20] Patel-Schneider, P.F., Hayes, P., Horrocks, I. OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 February 2004. Latest version available at: <http://www.w3.org/TR/owl-semantic/>
- [21] Radev, D.R., Jing, H., Budzikowska, M. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation and User Studies. In *Proc. of ANLP/NAACL 2000 Workshop*, (2000), 21–29
- [22] Sabidussi, G. The Centrality Index of a Graph. *Psychometrika*, 31 (1966), 581–603
- [23] Sheskin, D.J. Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press, (1997)
- [24] Storey, M.A., Musen, M., Silva, J., Best, C., Ernst, N., Ferguson, R., Noy, N.F. Jambalaya: Interactive Visualization to Enhance Ontology Authoring and Knowledge Acquisition in Protege. In *Proc. of Workshop on Interactive Tools for Knowledge Capture*, (2001)
- [25] Tummarello, G., Morbidoni, C., Puliti, P., Piazza, F. Signing Individual Fragments of an RDF Graph. In *Proc. of the 14th International Conference on World Wide Web (Special Interest Tracks and Posters)*, (2005), 1020–1021
- [26] Qu, Y. A Predicate-ordered Sort-ordered Logic for RDFS. In *Proc. of the 12th International Conference on World Wide Web (Poster Track)*, (2003)
- [27] Qu, Y., Hu, W., Cheng, G. Constructing Virtual Documents for Ontology Matching. In *Proc. of the 15th International Conference on World Wide Web Conference*, (2006), 23–31
- [28] Wang, T.D., and Parsia, B. CropCircles: Topology Sensitive Visualization of OWL Class Hierarchies. In *Proc. of the 5th International Conference on Semantic Web*, (2006), 695–708
- [29] White, D.R., and Borgatti, S.P. Betweenness Centrality Measures for Directed Graphs. *Social Networks*, 16 (1994), 335–346
- [30] Zhang, X., Li, H., Qu, Y. Finding Important Vocabulary within Ontology. In *Proc. of the 1st Asian Semantic Web Conference*, (2006), 106–112