RESEARCH ARTICLE OPEN ACCESS

Semantic WEB Services Using Clustering Approach

Jayeeta Majumder, Saikat Khanra
Department of Computer Science & Engineering &Technology Haldia Institute of Technology Haldia, West Bengal, India
Department of Computer Science Contai Polytechnic, Govt. of West Bengal Contai, West Bengal, India

Abstract:
Semantic Web Services, like conventional web services, are the server end of a client–server system for machine-to-machine interaction via the World Wide Web. Semantic services are a component of the semantic web because they use markup which makes data machine-readable in a detailed and sophisticated way (as compared with human-readable HTML which is usually not easily "understood" by computer programs). Semantic similarity measures are specific types of Semantic measures: mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a numerical description obtained according to the comparison of information formally or implicitly supporting their meaning or describing their nature.
Keywords: semantic web, Semantic Clustering, Irrelevant services.

I. Introduction

The Internet has been acknowledged as one of the recent technological revolutions, due to its large impact in the whole society. Nevertheless, precisely due to its impact, limitations of the current internet are becoming apparent; in particular, its inability to take into accounts in an automatic way the meaning of online documents. Some proposals for taking meaning into account began to appear, mainly the so-called ―"Semantic Web", which includes a set of technologies like Resource Description Framework (RDF)[2], based on new markup languages. Though these technologies could be technically sound, practical limitations, such as the high training level required to construct semantic web pages, and the small proportion of current semantic web pages –which, circularly produces low commercial interest in RDF, end up making the semantic web marginal today and also in the near foreseeable future. So, other options for taking into account semantics of today's internet were proposed. From ―latent‖ semantics [11] to fuzzy techniques and many others, they have in common their interest in modeling in an approximate way the meaning of online documents, at least for identifying their subject or topic. Extensive use of counting, statistical methods can bring to front many semantic "hidden" regularities of the web.

II. Semantic WEB

The Semantic Web[2][13] aims to add a machine tractable, re-purpose able layer to compliment the existing web of natural language hypertext. In order to realize this vision, the creation of semantic annotation, the linking of web pages to ontologies, and the creation, evolution and interrelation of ontologies must become automatic or semi-automatic processes. Semantic Web Services (SWSs) [15] go beyond current services by adding ontologies [17] and formal knowledge to support description, discovery, negotiation, mediation and composition. Finally, tools and infrastructures for the Semantic Web on the one hand and language technology on the other have so far remained largely independent from each other, despite the fact that they share a number of components, namely ontologies and reasoning mechanisms. HLT (Human Language Technology) systems can benefit from new developments like the Ontology Middleware Module.

III. Semantic similarities measure

Similarity measure provides a useful light weight approach [6] to exploit the available semantic metadata. In a large scale heterogeneous distributed environment (i.e the Grid), the computationally intensive process of logical reasoning can rarely be used to achieve a satisfactory result under time restraints Semantic similarity is an important concept that has been widely used [11] in many areas of research. The following is some approaches to semantic similarity measurement.

1 *Distance Metric for Semantic Nets*
2 *Information Based Measure*
3 *Similarity for Ontology Framework*

IV. Automatic topics discovery from hyperlinked documents

Topic discovery is an important means for marketing, e-Business and social science studies. As well, it can be applied to various purposes, such as identifying a group with certain properties and observing the emergence and diminishment of a certain cyber community. Automatic Topic Discovery (ATD) [14] method, which combines a method of base set construction, a clustering algorithm and an iterative principal eigenvector

computation method to discover the topics relevant to a given query without using manual examination. Given a query, ATD returns with topics associated with the query and top representative pages for each topic. An automatic topic discovery algorithm for a user given query, called ATD algorithm. ATD algorithm is composed of a method of base set construction, a clustering algorithm and a principal eigenvector computation algorithm. The aim of the ATD algorithm is to identify and isolate each strongly inter-connected cluster as topic in the web vicinity graph, and then select top-ranked web pages within each cluster to be its representing concept. The task of automatic topic discovery is composed of five fundamental parts:

- INPUT: a broad-topic query.
- A method to build a web vicinity graph for the query using either a focused crawler or a search engine to provide web pages related to the input query that then employs hyperlink expansion to create a web vicinity graph.
- A clustering algorithm to partition the web graph into separate clusters.
- A ranking algorithm to rank web pages within each cluster.
- OUTPUT: Each cluster is regarded as a topic and top-ranked web pages within the cluster are presented to the user as the representation of the topic.

4.1. Approximate use of hyperlinks:

Many files on a typical computer can be loosely divided into documents and data. Documents like mail messages, reports, and brochures are read by humans. Data, like calendars, address books, play lists, and spreadsheets are presented using an application program which lets them be viewed, searched and combined in many ways.

Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML), a markup convention that is used for coding a body of text interspersed with multimedia objects such as images and interactive forms. Metadata tags [7].

V. Semantic Clustering

Efficiently finding Web services [11] on the Web is a challenging issue in service-oriented computing. Currently, UDDI is a standard for publishing and discovery of Web services, and UDDI registries also provide keyword searches for Web services. However, the search functionality is very simple and fails to account for relationships between Web services. Firstly, users are overwhelmed by the huge number of irrelevant returned services. Secondly, the intentions of users and the semantics in Web services are ignored. Inspired by the success of partitioning approach used in the database design.

Clustering semantic approach (CSA) [16] is dependent on combination of the keyword technique and the semantics extracted from the services' descriptions. The objectives of CSA are to diminish the cost of computing a large dataset and to match services at the semantic concept level. The CSA approach is based on the assumption that the efficiency of finding services can be improved if irrelevant data can be eliminated before the extracting semantics algorithm is implemented.

In this section we propose our clustering probabilistic semantic approach (CPLSA) [10] for efficiently finding Web services. the samples returned may include irreverent services with respect to a query, so we first filter out those Web services whose contents are not compatible to a user's query to form a working dataset. Then we apply PLSA [10] to the working dataset for further clustering the dataset into a finite number of semantically related groups.

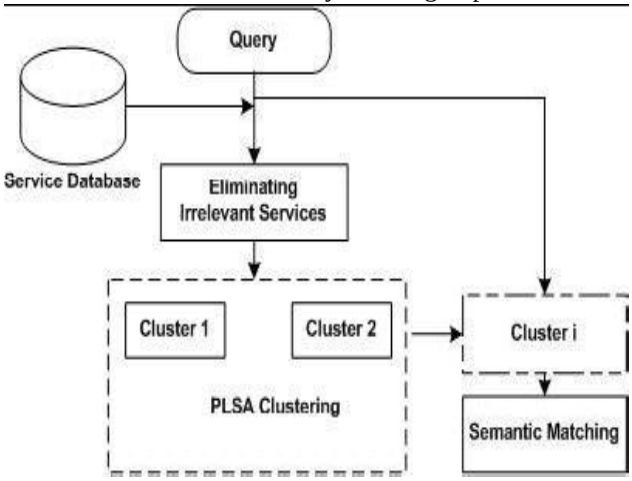

Figure 1 illustrates the outline of the proposed clustering semantic probabilistic approach (PLSA).

VI. Eliminating Irrelevant Services from Service Collection

We first retrieve a set of samples of Web services from a source Web services. Given a query q, a source of services would return a set of services based on some kind of similarity. To calculate the similarity, we use the Vector Space Model (VSM)[18] to represent Web services as points in syntactic space. Based on VSM, we can measure the similarity between a query q and a service s in the samples by computing the cosine of the angle between query vector q and service vector s as:

$$\text{Sim}(q\ s) = |q*s| / ||q||^2 * ||s||^2$$

Using the above similarity computation, we can acquire an initial set of samples of services through selecting a predefined threshold. Considering the possibility that the initial set of services may contain the services whose contents are not compatible with a user's query, we eliminate them accordingly from the sample set to improve the efficiency of service discovery, and also to reduce the cost of computation. Intuitively, these irrelevant data may have some negative impact on efficiently finding Web services; for one thing, the data may diminish the accuracy of the learning algorithms; for the other, they would

increase the computational load. Therefore, as the first step towards efficiently locating Web services, these irrelevant services should be eliminated before the clustering semantic algorithm is implemented. Several ways can be used to remove unrelated data from a dataset. One of the possible solutions is based on the feature selection, as indicated in [14]. This approach first sets a numerical threshold, and then computes the number of times a data object appears in a collection. If the number of times an object appearing in a collection is less than the predetermined threshold, the object is regarded as unrelated data and should be removed.

VII. Semantic algorithms scalability

RDF Growth, an algorithm that addresses a specific yet important scenario: large scale, end user targeted, metadata exchange P2P applications. In this scenario, peers perform browsing and querying of semantic web statements on a local database without directly generating network traffic or remote query execution. The database grows by learning from other peers in the P2P group using only a minimal amount of direct queries that are guaranteed to be executable with a low, predictable computational cost.

The principal component of the algorithm is the ―synchronize‖ procedure. Taking a URI as a parameter, the peer calls LOOKUP to receive a set of remote EPs. It removes the ones with the same signature as that calculated about the URI on the local DB and will call a heuristic *Hrdfn* which will suggest, using the signatures provided, the best remote EP to get information from. If a valid reply is received when requesting the RDFN from a remote EP, the peer will import the data into the local database. To keep the local EP and our ―public state‖ updated, the signature is then recalculated and the EP republished. Before republishing the EP, the peer checks if it is in possessions of information not otherwise known in the group, that is, if the newly calculated signature is not among those of the received EPs. This is usually the case when the peer synchronizes a URI about which new information was inserted locally. If this is the case, it will attempt to "broadcast" or, if not available, issue a "newsflash" procedure before reinserting. If, at the earlier stage, the GETRDFN had failed, the peer would have removed the corresponding EP from the set and proceeded in the loop. As a result of this procedure, at the end of the transient period, the local RDFN about a URI will converge to the one of the other peers that also chose to publish and synchronize.

VIII. Semantics-leveraged search

Historically, information retrieval (IR) [17] has followed two principally different paths that we call syntactic IR and semantic IR. In syntactic IR, terms are represented as arbitrary sequences of characters and IR is performed through the computation of string similarity. In semantic IR, instead, terms are represented as concepts and IR is performed through the computation of semantic relatedness between concepts. Semantic IR, in general, demonstrates lower recall and higher precision than syntactic IR. However, so far the latter has definitely been the winner in practical applications.

IX. Applications of approximate semantics

The idea of a *semantic web*, able to describe, and associate meaning with data, necessarily involves more than simple XHTML mark-up code [11]. It is based on an assumption that, in order for it to be possible to endow machines with an ability to accurately interpret web homed content, far more than the mere ordered relationships involving letters and words is necessary as underlying infrastructure, (attendant to semantic issues). Otherwise, most of the supportive functionality would have been available in Web 2.0 (and before), and it would have been possible to derive a semantically capable Web with minor, incremental additions.

Additions to the infrastructure to support semantic functionality include latent dynamic network models that can, under certain conditions, be 'trained' to appropriately 'learn' meaning based on order data, in the process 'learning' relationships with order (a kind of rudimentary working grammar)

X. Conclusion

Semantic Web Technology Today gives --- World Wide Web incremental advance

- Evolvable approach to information
- Leverages open software building blocks
- Builds on diversity
  – creating new knowledge
  – enabling new applications
- Low-risk adoption strategy

References

[1] W. Abramowicz, K. Haniewicz, M. Kaczmarek and D. Zyskowski. Architecture for Web services filtering and clustering. In Internet and Web Applications and Services, (ICIW '07), 2007.

[2] C. Atkinson, P. Bostan, O. Hummel and D. Stoll. A Practical Approach to Web service Discovery and Retrieval. In 2007 IEEE International Conference on Web services (ICWS 2007),2007

[3] M. W. Berry, S. A. Pulatova and G. W. Stewart. Computing Sparse Reduced-Rank Approximations to Sparse Matrices. In ACM Transactions on Mathematical Software, Vol. 31, No. 2, Pages 252–269, 2005.

[4] J. Baliński and C. Daniłowicz. Re-ranking Method based on Inter-document Distances. In Journal of the Information Processing and Management.. V. 41, Issue 4, 2005.

[5] I. Constantinescu, W. Binder and B. Faltings. Flexible and efficient matchmaking and ranking in service directories. In Proceedings of the IEEE International Conference on Web Services (ICWS'05), 2005.

[6] X. Dong, A. Halevy, J. Madhavan, E. Nemes and J. Zhang. Similarity Search for Web services. In Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[7] J. T. Giles, L. Wo and M.W. Berry. GTP(General Text Parser) software for text mining. In Statistical Data Mining and Knowledge Discovery, H. Bozdogan, ed., CRC Press, Boca Raton, FL, 2003, papers: 455-471. 2003

[8] J. Garofalakis, Y. Panagis, E. Sakkopoulo and A. Tsakalidis. Web service Discovery Mechanisms: Looking for a Needle in a Haystack? In International Workshop on We b Engineering, August 10, 2004.

[9] A. Hess and N. Kushmerick. Learning to Attach Semantic Metadata to Web services. In 2nd International Semantic We Conference (ISWC2003), Sanibel Island, Florida, USA, 2003

[10] T. Hofmann. Probabilistic Latent Semantic Analysis. In Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval Berkeley, California, pages: 50-57, ACM Press, August, 1999.

[11] H. Lausen and T. Haselwanter. Finding Web services. In the 1st European Semantic Technology Conference, Vienna, Austria,2007

[12] M. Klein and A. Bernstein. Toward High-Precision Service Retrieval. In IEEE Internet Computing, Volume: 8, No. 1, Jan.– Feb. pages: 30 – 36, 2004.

[13] J. Ma, J. Cao and Y. Zhang. A Probabilistic Semantic Approach for Discovering Web services. In The 16th International World Wide Web Conference (WWW2007). Banff, Alberta, Canada, May 8 -12, 2007.

[14] B. Mandhani, S. Joshi and K. Kmmamuru. A Matrix Density Based Algorithm to Hierarchically Co-Cluster Documents and Words. In the 12th International World Wide Web Conference (WWW2003). May 20- 24, Budapest, Hungary, 2003.

[15] M. Paolucci, T. Kawamura, T. Payne and K. Sycara. Semantic Matching of Web services Capabilities. In Proceedings of the 1st International Semantic Web Conference (ISWC2002). 2002.

[16] R. Nayak and B. Lee. Web service Discovery with Additional Semantics and Clustering. In Web Intelligence, IEEE/WIC/ACM International Conference, 2007

[17] K. Sivashanmugam, K. Verma, A.P and J.A. Miller. Adding Semantics to Web services Standards. In Proceedings of the International Conference on Web services ICWS'03, pages: 395-401, 2003.

[18] G. Salton. Automatic Text Processing—The Transformation, Analysis, and Retrieval of Information by Computer. In Published by Addison-Wesley Publishing Company. 1988.

[19] A. Sajjanhar, J. Hou and Y. Zhang. Algorithm for Web services Matching. In Proceedings of the 6th Asia- Pacific Web Conference, APWeb 2004, angzhou, China, April 14-17, 2004, Lecture Notes in Computer Science 3007 springer 2004.

[21] XMethods. http://www.xmethods.com/

[22] http://www.census.gov/epcd/www/naics. html.

[23] http://www.Webservicelist.com