Proceedings of the 4th National Conference; INDIACom-2010
Computing For Nation Development, February 25 – 26, 2010
Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

An Efficient Knowledge Base Mutual Linker Framework

Syed Mohiuddin
Assistant Professor, CSE, Muffakham Jah College of Engineering & Technology
syed@mjcollege.ac.in

Himayatullah Sharief
Assistant Professor, CSE, Muffakham Jah College of Engineering & Technology
shariffmca@gmail.com

Dr. T. Eshwar
Professor, CSE, Muffakham Jah College of Engineering & Technology
etenneti@gmail.com

ABSTRACT
In this paper, we introduce a generalization of the automatic linking engine which can automates the process of linking disparate data entries into a fully connected conceptual network. The main challenges of this problem include correctly identifying which terms to link and which entry to link to with minimal effort on the part of users, efficiency and scalability of the links, and simplification to multiple knowledge bases and web-based information environment. We present the approach that utilizes subject classification and other metadata to address these challenges. We also present evaluation results demonstrating the effectiveness and efficiency of the approach.

KEYWORDS
E-learning, automatic linking, wiki, Semantic Web.

INTRODUCTION

Collaborative online knowledge bases such as encyclopedias are becoming increasingly popular because of their open access, comprehensive and interlinked content, rapid and continual updates, and community interactivity. To understand a particular concept in these knowledge bases, a reader needs to learn about related and underlying concepts. Thus, it is critical that users of any online reference are able to easily access requisite concepts in the network in order to fully understand. For full comprehension, these accesses should extend all the way to the concepts that are evident to the reader's perception.

The popularity of these encyclopedic knowledge bases has also brought about a situation where the availability of high-quality, canonical definitions and declarations of educationally useful concepts have outpaced their usage in other educational information resources on the web. Instead, the user must execute a new search to look up an unknown term when it is encountered, if it is not linked to a definition. For example, blogs, research repositories, and digital libraries quite often do not link to definitions of the concepts contained in their texts and metadata, even when such definitions are available. This is generally not done because of the lack of appropriate software infrastructure and the extra work creating manual links required. When such linking is actually done, it tends to be incomplete and is quite painstaking.

PROBLEM DEFINITION

We study the problem of invocation linking in this paper to build a semantic network for collaborative online encyclopedia. We first define a number of terminologies and define our problem to facilitate our discussion. For our purpose, a collaborative online encyclopedia is a kind of knowledge base containing standardized knowledge contributed by a large number of participants. Any article submitted by a user in such a collaborative amount is an entry or an object. We say invocation referring to a specific kind of semantic link that of concept invocation. Concept invocation refers to the use of a concept in the text. Any statement in a language is composed of concepts represented by tuples of words. Such a statement invokes these concepts, as evidenced by the inclusion of word tuples that correspond to common labels for the concepts. We call these tuples of words concept labels. An invocation link is a hyperlink from these tuples of words in an entry that represent a concept to an entry that defines the concept. We refer to the tuples of words being linked from as link source and the entry being linked to as link target. The problem of invocation linking is how to add these invocation links in a collaborative online encyclopedia. While it is possible to extend the problem definition and the techniques developed in this paper for other types of linking such as links to articles with a similar or different point of view, it is our focus in this paper to study concept or definitional linking.

EXISTING SOLUTIONS

We briefly survey the existing and potential solutions for invocation linking and motivate our automatic linking

approach. The existing and potential linking approaches can be mainly classified into the following three categories, namely, manual linking, semiautomatic linking, and automatic linking.

Manual linking refers to the linking technique where both the link source and link target are explicitly defined, e.g., anchor tags in html documents. Most current online encyclopedias use the manual approach. Even the Blog software generally requires writers create links manually.

Semiautomatic linking refers to the technique where the terms at the source are explicitly marked for linking, but the link target is determined by the mutual online encyclopedia system. Some current online encyclopedias use the semiautomatic approach.

Automatic linking or more specifically automatic invocation linking refers to the technique where the terms at the source and link target are both automatically determined by the system. This is the approach that we believer in order to build the semantic network with minimal manual effort.

Encyclopedias use a semiautomatic approach. That is, the links are manually delimited by authors when the author invokes a concept that they believe should be defined in the collection, but the system figures out the destination. If an entry for a concept is present only by an alternate name, the link might fail to be connected. Links to nonexistent entries are rendered specially as broken links. However, this is inherently somewhat distracting to those uninterested in creating a new entry. The other systems that take a similar approach also fail to provide systemic treatment of homonymy. The semi automatic convention is to manually create disambiguation nodes, which contain links to all homonymous concepts with a particular label. Such nodes add an extra step to navigation, require ongoing maintenance, and can contain an extremely random and distractive jumble of topics.

The perception taken in our work is that the manual and semiautomatic approaches are an unnecessary burden on contributors, since the knowledge management environment should know which concepts are present and how they should be cited. By contrast, authors will usually not be aware of all concepts that are already present within the system especially for large or distributed knowledge base network. A more challenging problem with the manual and semiautomatic linking strategy is that a growing, dynamic knowledge base will generally necessitate links from old entries to new entries as the collection becomes more complete. To attend to this reality would require continuous re-inspection of the entire knowledge base by writers or other maintainers, which is very complex problem. To keep an evolving knowledge base fully linked, it would be necessary for maintainers to search it upon each update or periodically to determine if the links in the constituent articles should be updated. When generalizing to inter-linkage across separate knowledge base, the task would potentially be even more laborious, as authors would have to search across multiple web sites to determine what new terms are available for linking into their entries. There are a number of potential technologies that one might apply to the automatic linking approach. We briefly review them below and discuss their limitations and implications on the automatic linking problem.

One technology is the information retrieval approach for web search. One part of our problem in identifying the best linking target for a concept label bears similarity to the search problem in finding the most relevant documents based on a keyword. Yet, for the most part, the work in IR has not been explored in the collaborative semantic linking context. While typical IR issues such as plurality, homonyms, and polysemy are all relevant for the linking process, some of the techniques are not directly applicable. For example, the traditional IR approach relies on term-frequency and inverse document-frequency to rank the retrieved documents. In our problem, the entries that define a particular concept may not contain the actual concept label. Thus, the term-frequency based approaches are not directly applicable in identifying the best link target or entry. In addition, in our problem, not only the link target but also the link sources need to be identified and linked automatically.

Another technology is the recommender systems that aim to predict ratings of a particular item for a particular user using a set of similar users based on a user-item rating matrix. At an initial glance, we can model our problem as an entry-entry link matrix where each cell represent a link or non-link from a certain entry to another entry and use entry similarities to help determine the best entry to link to for a term that belongs to a certain entry. While this approach is more appropriate for relevance linking and may help to narrow down the potential link targets, it alone is not sufficient for the invocation or concept linking problem. Nevertheless, it is on our research agenda to enhance our current link ranking strategy by adapting the collaborative filtering technologies to enhance the linking precision by incorporating entry similarities and user feedback into the linking process.

DESIGN GOALS

The optimal end product of an automatic invocation linking system should be a fully connected network of articles that will enable readers to navigate and learn from the corpus almost as naturally as if was interlinked by painstaking manual effort. Without understanding the invoked concepts in a statement, the reader cannot attain a complete understanding of the statement, and by extension the entry it appears in. This is why inter-linkage is so important in hypertexts being used as knowledge bases, and why we believe an automated system is of such utility. In this paper, we sought to build an automatic linking system using the metadata of the entries. Building such an automatic invocation linking system for a collaborative online encyclopedia presents a number of computing challenges. We

discuss the challenges below and outline our design goals to address them.

The first challenge addressed is Linking quality. The main analytic challenges lie in how to determine which terms or phrases to link and which entries to link to. Typical IR and natural language processing issues such as plurality, homonyms, and polysemy are all relevant for the linking process and bear on the quality of linking. In light of all these challenges, the linking process is necessarily imperfect and so linking errors may be present. We characterize many such forms of errors as follows:

Mislinking: refers to the error that a term or phrase is linked to an incorrect link target.
Overlinking: refers to the error that a term or phrase is linked when there should be no link at all. Note that Overlinking also contributes to mislinking because the term is mislinked.
Underlinking: refers to the error that a term or phrase is not linked when there should be a link because it invokes a concept that is defined in the knowledge base.

An important goal of designing the automatic linking system is to reduce the above errors and improve the link precision perfect link precision means every link is linked to the correct link target while maintaining high link recall perfect link recall means a link is created for every concept label that can and should be linked given the present state of the knowledge base.

The second challenge is Efficiency and scalability. Another important design goal of an automatic linking system is its efficiency so the links can be created near-real time during rendering of the entries and its scalability so it can handle the large size of an online encyclopedia corpus. In addition, most collaborative corpora change frequently, an automatic invocation linking system needs to efficiently update the links between entries that are related to newly defined or modified concepts in the knowledge base. A continually changing knowledge base must be dealt in such a way that the analysis and processing of automatic links is tractable and scalable.

The third challenge is Generalization to multiple knowledge base. It is also necessary and important that an automatic linking system is easy to use for the adoption by a large user base and easy to set up for the widespread adoption for linking various materials across multiple sites.

To help users learn more quickly, it is now generally accepted that knowledge bases should leverage each others content or metadata to increase the scope of the available learning materials. Our design goal is to leverage these standards so that our automatic linking system would not only enable intra-linking collaborative encyclopedias but also allow for linking educational materials such as lecture notes, blogs, and abstracts in research and educational digital libraries. Such usage could aid researchers and students in the better understanding of abstracts and full texts, and could also help them find related articles quickly.

SYSTEM FRAMEWORK

In this section, we present the model behind our system and discuss key techniques and features in the framework. In this system when an entry is rendered either at display time or during offline batch processing, the text is scanned for words or concept labels or link source and they are ultimately turned into hyperlinks to the corresponding entries or link target in the output rendering. There are two basic steps in performing the invocation linking. When an article is submitted, the system starts link source identification by pulling out un-linkable portions of text that need to be escaped and replaces them by special tokens. The engine then breaks the text of an entry into a single words/tokens array. The tokens and token phrases or sentences that invoke concepts defined in other entries are then used for link target identification to determine the entries to link to. The system architecture is depicted in Figure 1.

In order to determine which entry to link to for a concept label, our system indexes the entries by building a concept map that maps all of the concept labels in the corpus to the entries which define these concepts. The tokens and token phrases that are identified as link sources are searched to retrieve the candidate links using the concept map. After the candidate links are determined, they are filtered based on linking policies. The candidates are then compared by nearness and the object with the closest classification is then selected as the link target.

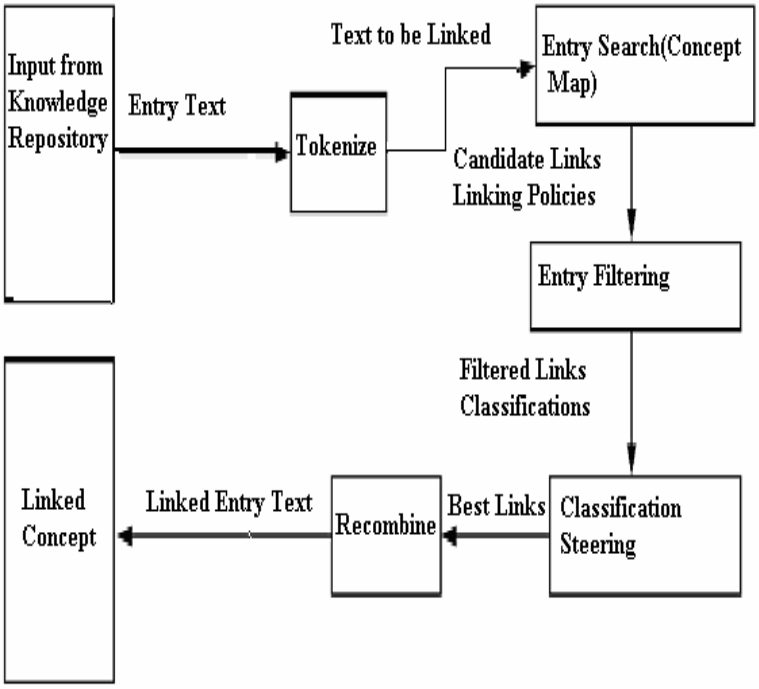

Figure 1: System Architecture

The closest candidate for each position is then substituted into the original text and the linked document is then returned. Figure below illustrates the conceptual flow of the automatic linking process. In addition, when new concepts are added to the collection or the set of concept labels otherwise changes, entries containing potential invocation of these concept labels can be invalidated. This allows entries to be rescanned for links, either at invalidation time or before the next time they are displayed. Our Systems uses a special structure called the invalidation index to facilitate this. This automatic system almost completely frees content authors from having to bother about links. It addresses the problems of both outgoing and incoming links, with respect to a new entry or new concepts. However, it is not completely infallible, and in an epistemological sense, there is only so much that a system can infer without having a human-level understanding of the content. Because of this, the user can ultimately override the automatic linking, create their own manual links, or specify link policies for steering the automatic linker. While complemented and enhanced by the interactive learning components, our system is a completely automatic system, and performs well even without any human efforts.

IMPLEMENTATION

The core methods of our system have essentially proven their large-scale applicability in the tutorials site, which contains large number of entries defining more concepts. Our system is developed with java and is designed to have the minimum amount of dependencies necessary while still running efficiently. Thus, our system only requires a database system and corresponding XML packages. Our system has been designed in such a way that it can be used with any document knowledge base. One of the design goals of our system was ease of deployability, portability and use. For this reason, our system uses Java programming Language and simple XML formats for its communications. All communications in our system are over socket connections, and all requests and responses with the server are in XML format.

CONCLUSION

We have presented an automatic linking system for providing invocation linking capabilities for online collaborative encyclopedia. We outlined the design goals for any automatic linking system should strive to achieve. We presented a set of experiments demonstrating the effectiveness and efficiency of our approach. Our work continues along several threads. First, we are exploring automatic keyword extraction techniques to better extract concept labels to be linked. Second, we are exploring reputation systems and collaborative filtering techniques to further enhance the link navigation by addressing issues of competing entries and different needs and preferences of people. This especially becomes an issue when one goes beyond a single collaborative knowledge base, as would typically be the case in linking to them by third parties. Finally, a major research and development item is the generalizing for interlinking of multiple knowledge bases across domains with expansion of knowledge base systems.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, June 2005.

[2] R.A. Baeza-Yates and B.A. Ribeiro-Neto, Modern Information Retrieval. ACM Press/Addison-Wesley, 1999.

[3] K. Cardinaels, M. Meire, and E. Duval, “Automating Metadata Generation: The Simple Indexing Interface,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05), pp. 548-556, 2005.

[4] A. Souzis, “Building a Semantic Wiki,” IEEE Intelligent Systems, vol. 20, no. 5, pp. 87-91, Sept./Oct. 2005.