NNexus: Towards an automatic linker for a massively-distributed collaborative corpus

James Gardner [*]
Department of Math&CS
Emory University
jgardn3@emory.edu

Aaron Krowne
Woodruff Library
Emory University;
PlanetMath.org
akrowne@emory.edu

Li Xiong
Department of Math&CS
Emory University
lxiong@mathcs.emory.edu

Abstract

Collaborative online encyclopedias such as Wikipedia and PlanetMath are becoming increasingly popular. In order to understand an article in a corpus a user must understand the related and underlying concepts through linked articles. In this paper, we introduce NNexus, a generalization of the automatic linking component of PlanetMath.org and the first system that automates the process of linking encyclopedia entries into a semantic network of concepts. We discuss the challenges, present the conceptual models as well as specific mechanisms of NNexus system, and discuss some of our ongoing and completed works.

1 Introduction

Collaborative online encyclopedias or knowledge bases such as Wikipedia[1] and PlanetMath[2] are becoming increasingly popular because of their open access, comprehensive and interlinked content, continual updates, and community interactivity.

Motivation. Because knowledge can be modeled as taking the shape of a sematic network, we view an online encyclopedia as a set of concepts whose meanings depend on the text of entries and on their position in the semantic network (the connections to other entries). To understand a particular concept, a reader needs to learn about related and underlying concepts. It is of paramount importance that users of any online reference are able to "jump" to requisite concepts in the network in order to understand the current one—all the way down to the concepts that are so simple they are evident to the reader's intuition.

Most current online encyclopedias (including Wikipedia) require the author of an article to explicitly create links to other articles in order to build this link network. However, this task is an unnecessary burden on users, since the right database should "know" which concepts are present. By contrast, authors will usually not be aware of all requisite concepts which are already present within the system—especially for a large database. Even more challenging, a growing collection 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 corpus by writers or other overseers.

To ameliorate these problems with minimal manual effort, *automatic invocation linking* between entries in the online corpus is needed. The end result of any such system should be a semantic 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.

Challenges. Building an automatic linking system for a collaborative online encyclopedia presents a

[*] This work was supported by the Google Summer of Code program, summer, 2006.

[1] http://www.wikipedia.org

[2] http://www.planetmath.org

number of research challenges. The main challenges lie in how to determine which terms or phrases to link and which entries to link to. Typical information retrival and natural language processing issues such as plurality, homonyms, and polysemy all affect the quality of linking and can contribute to linking errors. Some of these errors take the form of links citing the incorrect homonym from a group of homonyms, while some take the form of linking when there should be no linking at all—a phenomenon which we call specifically *overlinking*. We use *mislinking* as a term to refer to any type of reduced *link precision* (the fraction of created links which are correct). An important goal of an automatic linking system is to improve the linking precision while maintaining high link recall (a link created for every concept label that can and should be linked given the present corpus).

There are many standard methods for improving searching quality in IR literature and current search engines [2] and they have not been explored in the collaborative semantic linking context [3]. In addition, online encyclopedias are typically organized into a classification hierarchy, and we argue that this knowledge can be utilized in order to dramatically increase the precision of automatic linking, and largely solve the polysemy problem.

Contributions. Bearing these issues in mind, we designed and developed NNexus (Noosphere Networked Entry eXtension and Unification System), a system used to automate the process of automatically linking encyclopedia entries (or other definitional knowledge bases) into a semantic network of concepts. NNexus has application to digital libraries or other types of online references or knowledge bases.

NNexus is an abstraction and generalization of the automatic linking component of the Noosphere system [4], which is the platform of PlanetMath (planetmath.org), PlanetPhysics (planetphysics.org), and other Noosphere sites. To the best of our knowledge, it is the first automatic linking system that links articles and concepts with the use of a classification scheme, to make linking almost a "non-issue" for writers, and completely transparent to readers.

The key features of NNexus include a customized information-retrieval based automatic linking system and a set of techniques such as ontology/subject driven link steering, and declarative linking priorities and clauses that are specifically designed to enhance the linking precision for a minority of "tough cases." We are also researching the ability to add reputation based or collaborative filtering techniques to link steering, given the bearing of author personalization on linking.

The rest of the paper has been constituted so that a reader (including users and developers) can understand the conceptual models and theories behind NNexus while also learning some specific mechanisms behind its operation.

2 NNexus System

Users of NNexus will apply the following basic functionality to their corpus. When an entry is rendered (either at display time or during offline batch processing), the text is broken down into tokens and scanned for words that invoke concepts that have been defined in other entries. These words (or word tuples) are ultimately turned into hyperlinks to the corresponding entries in the output rendering. In addition, when the concepts are added to the collection (or the set of concept labels otherwise changes), entries containing potential invocation of these concept labels are *invalidated* using a special inverted index called the *invalidation index*. This forces these entries to go through link analysis themselves by or before the next time they are displayed.

This automatic system almost completely frees the author from having to "think 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 a theoretical sense, there is only so much that a system can infer without having a human's level of understanding of the content. Because of this, the user can ultimately override the automatic linking, or create their own original linking.

NNexus was developed with Perl and was designed to have the minimum amount of dependencies necessary while still running efficiently. Thus, NNexus only requires a database system (currently MySQL is supported) and some Perl XML packages (available

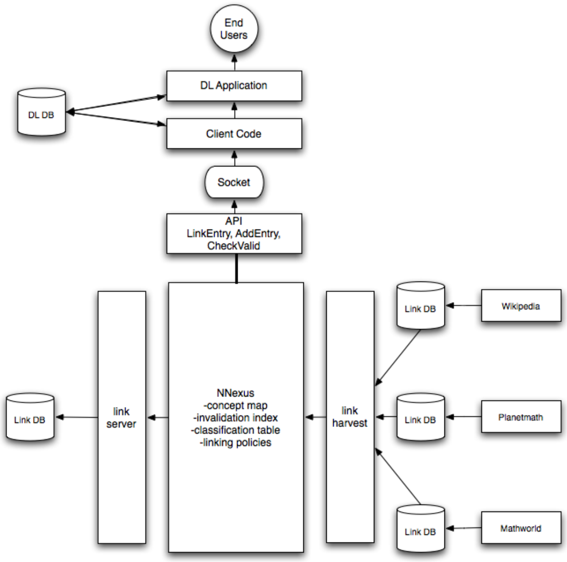

Figure 1: NNexus System Architecture

from CPAN). NNexus has been designed with an API so that it can be used with any document corpus. Figure 1 shows a sketch of NNexus system architecture. Due to space limitations, we omit the implementation details and instead present an overview of algorithms and a number of key features of NNexus system in this section.

2.1 Overview of Techniques

Indexing. NNexus 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. To facilitate efficient scanning of entry text to find concept labels, the map is structured as a chained hash, keyed by the first word of each phrase placed in it.

Entry Search. When an article is submitted, NNexus breaks the text of an entry into a single words/tokens array to iterate through. If a word matches the start of an indexed concept label, the following words in the text are checked to see if they match the longest concept label starting with that word. If this fails, the next longest concept label is checked, and so on. When a matching concept label is found, it is included in the *match array.*

Entry Filtering and Selection. After the match array is built, each match is checked against the *linking policies* (see Section 2.2) for the target articles. If the match is forbidden from linking to the target it will be removed from the match array. The match subroutine also assigns a priority value to each match to determine which article it should link to. We also have a few efforts under development that explore various ranking techniques by integrating multiple factors such as domain class, priority, and reputation of the entries.

Linking. Finally, the linked text array is recombined with the removed tokens, and returned as a single text string. This final string is returned to the calling user program.

2.2 System Features

Longest phrase match. NNexus always performs longest phrase match. For example, if the writer mentions the phrase “green widget” in their “thingamagig” entry, and there is not only a “widget” entry, but also a “green widget” entry in the collection, NNexus links to the “green widget” entry.

Morphology. NNexus also performs some morphological transformations on concept labels in order to ensure they can be linked to. The first, and most important transformation, has the effect of invariance of pluralization. The second invariance is due to possessiveness. Another morphological invariance concerns international characters. The final transformation has the effect of invariance under indexing style. When a token is checked in the index NNexus will ensure that the token is singular and non-possessive.

Link Suppression. Often automatic reference linking may be a little overzealous. A user may use a word in a natural language sense (e.g. “even”) which is also the title for an encyclopedia object (e.g. “even number”) due to the ambiguity in natural language and names of mathematical concepts. For this reason, users can escape certain words and phrases from being linked by NNexus or override the automatic linking.

Linking Policy. Central to expressing linking restrictions is the linking policy, a set of directives controlling linking based on the subject classification system within the encyclopedia. NNexus allows authors to permit or forbid certain classes of articles from linking into their articles. The linking policy of an article describes, in terms of subject categories, to where links may be made or prohibited. For example, the linking policy for an entry on group theory might simply be that terms in the entry can only be linked to if the linking object is in the "group theory" category. Alternatively, an entry on set theory (because it is so elementary) might allow everyone to link to the terms it defines *except* restrict articles in the image processing category from linking to the term "image" (the word "image" has different meanings in both areas).

3 Results and Discussion

We performed a mislinking and overlinking study in June 2006 on the planetmath collection which uses the basic system features of NNexus (no linking policies). About 12% of links were mislinks, 7.9% of links were overlinks, and 61.1% of the mislinks were overlinks. After applying the new linking policies we discovered that a considerable amount of mislinking was resolved, but a formal study has yet to be conducted. We estimate that about 95% of the links will be correct once the linking policies are imposed on most of the planetmath corpus.

A comparable system to Noosphere is Wikipedia. Wikipedia does not use automatic linking and thus has near-perfect linking precision. Links are manually-delimited by authors when the author invokes a concept that they believe should be in the collection. A survey in [5] shows that about 97-99% of wikipedia links are accurate, but is not fully comparable to our survey because it relies on disambiguation nodes and doesn't measure underlinking (imperfect link recall). Further, no formal comparison of the effort required to maintain entry interlinks has been made.

4 Future Directions

Our work in NNexus continues along several threads. In addition to improving the policy-based link selection, we are also exploring reputation systems and collaborative filtering techniques [1] to address issues of "competing" entries and different needs and preferences of authors.

We are also working towards enabling NNexus to allow the extension of the semantic network across multiple domains (e.g., when processing a mathematics article for determining links, the terms and phrases in the article may link to articles from sites such as Planetmath.org, Wikipedia.org, and MathWorld.com, etc.).

5 Conclusion

We have presented the challenges of automatically inter-linking a dynamic corpus and introduced NNexus, a modular system for performing this task. The achievements of the precursor to the NNexus system, the Noosphere automatic linker, can be seen at PlanetMath.[3] NNexus will soon be available for general use as open source software, and we will continue pursuing more powerful enhancements to the link-selection logic.

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 transactions on knowledge and data engineering*, 17(6), 2005.

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

[3] J. Kolbitsch and H. Maurer. Community building around encyclopeadic knowledge. *Journal of Computing and Information Technology*, 14, 2006.

[4] Aaron Krowne. An architecture for collaborative math and science digital libraries. Master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2003.

[5] G. Weaver, B. Strickland, and G. Crane. Quantifying the accuracy of relational statements in wikipedia: a methodology. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006.

[3] `http://planetmath.org/`.