WordNet-based Summarization to Enhance Learning Interaction Tutoring

Antonella Carbonaro
Department of Computer Science, University of Bologna,
antonella.carbonaro@unibo.it

Abstract
The process of summarizing information is becoming increasingly important in the light of recent advances in resource creation/distribution technology and the resulting influx of large amounts of information in everyday life. These advances are also challenging educational institutions to adopt the opportunities of distributed knowledge sharing and communication. The paper describes a summarization system to support tutors in managing student communication and interaction within a learning framework. Results show the adequacy of the system in identifying a good content summarization and then in improving the efficiency and effectiveness of the context in which summarization can be integrated.
1 Introduction

The Internet has grown beyond merely hosting and displaying information passively. It provides easy access for people to share, socialize, and interact with one another. Information displayed and exchanged between people are dynamic, in contrast to static information depicted in the older age of the Internet, such as that exchanged in forums, web virtual spaces where people can participate in discussions. The availability of vast amounts of thread discussions in forums has promoted an increasing interest in knowledge acquisition and summarization for forum threads.

Text summarization has been an interesting and active research area since the ‘60s. The definition and assumption is that a small portion or several keywords of the original document can informatively and/or indicatively represent it in its entirety. Reading or processing this shorter version of the document would save time and other computational resources (Zhou & Hovy, 2006). This characteristic is especially true and urgently needed at present due to the vast availability of information.

Moreover, the Web is becoming a social place and producing new applications with surprising regularity: there has been a shift from just existing on the Web to participating on the Web. Community applications and online social networks have become very popular recently, both in personal/social and professional/organizational domains (Kolbitsch & Maurer, 2006). Most of these collaborative applications provide common features such as content creation and sharing, content-based tools for discussions, user-to-user connections and networks of users sharing common interests, reflecting today’s Web 2.0 rich Internet application-development methodologies. The development of concept-based systems to facilitate knowledge representation and extraction and content integration is provoking a great deal of interest (Bighini & Carbonaro, 2004).

Concept-based approaches to represent dynamic and unstructured information can be useful to address issues like trying to determine the key concepts and to summarize the information exchanged within a personalized environment. Indeed, a virtual learning system is not just a set of contents, it may include collaboration spaces and tools such as forums, chats or shared document areas.

2 Interaction

The amount of interaction in technology-enhanced learning systems appears to be an important element of learning effectiveness. Wagner (1994) defined interaction as an interplay and exchange in which individuals and groups influence each other. Thus interaction focuses on the interpersonal behaviors
in a learning community. Gunawardena and Zittle (1997) argued that on-line
students can create social presence by projecting their identities and building
on-line communities through text-based communications alone.

In Rovai and Barnum (2003) also provided evidence that students’ perceived
that learning from on-line courses was positively related to quantitative
measures of course interaction. However, judgements about the relative impor-
tance of the two interaction variables are difficult because these variables are
correlated. Nonetheless, only the active interaction measure, representing, for
example, the number of student messages posted to discussion boards or the
number of participants in forum threads, was significant. This finding affirms
the importance of providing opportunities for on-line students to learn by active
interaction with each other and with the instructor (Zirkin & Sumler, 1995).
Consequently, educators should develop and include highly interactive material
in distance learning and encourage students to participate in on-line discus-
sions. Findings also suggest that passive interaction, analogous to listening to
but not participating in discussions, was not a significant predictor of percep-
ted learning in the present study. Consequently, using strategies that promote
active interaction appears to lead to greater perceived learning and may result
in higher levels of learner satisfaction with the on-line learning environment.
The quality of the interactions is another important aspect of communication
that should be a topic of further research.

A key issue in the use of learning systems is that tutors should be supported
in order to manage the communication facilities provided by the community
and to monitor student interactions. This aspect has been largely neglected in
the literature. However, supporting tutors is very important to make learning
communities effective.

Although some platforms offer reporting tools, when there are a great num-
ber of students and a great diversity of interactions, it becomes hard for a tutor
to extract useful information. Conceptual-based techniques can build analytic
models and uncover useful information from data.

The system we want to propose can find application in any context in which
the group interaction is a requisite, and we believe that a Web-based learning
system is an ideal application domain.

3 Summarization

Summarization is a widely researched problem. As a result, researchers have
reported a rich collection of approaches for document summarization.

There are two main types of approaches available in the literature on the
topic. The first is a class of approaches that deals with the problem of document
classification from a theoretical point of view, making no assumption on the
application of these approaches. These include statistical (McKeown et al., 2001), analytical (Brunn et al., 2001), information retrieval (Aho et al., 1997) and information fusion (Barzilay et al., 1999) approaches. The second class of resources deals with techniques that focus on specific applications, such as baseball program summaries (Yong Rui et al., 2000), clinical data visualization (Shahar & Cheng, 1998) and web browsing on handheld devices (Rahman et al., 2001). In addition, (NIST website) it reports a comprehensive review.

In this paper, a practical approach is proposed for extracting the most relevant keywords from the forum threads to form a summary without assumption on the application domain. The idea of our approach is to find out concepts from the keyword extraction based on statistics and synset extraction using WordNet. Then a semantic similarity analysis is carried out between the keywords to produce a set of semantic relevant words summarizing actual significance of the forum.

WordNet (Fellbaum, 1998) is an online lexical reference system, in which English nouns, verbs, adjectives and adverbs are organized in sets of synonyms. Each synset represents one sense, that is one underlying lexical concept. Different relations link the sets of synonyms, such as IS-A for verbs and nouns, IS-PART-OF for nouns, etc. Verb and noun senses are organized in hierarchies forming a “forest” of trees. For each keyword in WordNet, we can have a set of senses and, in the case of nouns and verbs, a generalization path from each sense to the root sense of the hierarchy. WordNet could be used as a useful resource with respect to the semantic tagging process and has so far been used in various applications including Information Retrieval, Word Sense Disambiguation, Text and Document Classification and many others.

Noun synsets are related to each other through hypernymy (generalization), hyponymy (specialization), holonymy (whole of) and meronymy (part of) relations. Of these, (hypernymy, hyponymy) and (meronymy, holonymy) are complementary pairs. The verb and adjective synsets are very sparsely connected with each other. No relation is available between noun and verb synsets. However, 4500 adjective synsets are related to noun synsets with pertainyms (pertaining to) and attra (attributed with) relations.

To extract important information from forum threads, we use the following feature extraction pre-process. Firstly, we label the occurrences of each word in the document as a part of speech (POS) in grammar. This POS tagger discriminates the POS in grammar of each word in a sentence. After labelling all the words, we select those labelled as nouns and verbs as our candidates. We then use the stemmer to reduce variants of the same root word to a common concept and filter the stop words.

A vocabulary problem exists when a term is present in several concepts; determining the correct concept for an ambiguous word is difficult, as is deci-
The concept of a document containing several ambiguous terms. To handle the word sense disambiguation problem we intend to use similarity measures based on WordNet.

The use of the described Word Sense Disambiguation step reduces classification errors due to ambiguous words, so to allow a better precision in the summarization process. For example, if the terms “procedure”, “subprogram” and “routine” appear in the same resource, we consider three occurrences of the same synset “{06494814}: routine, subroutine, subprogram, procedure, function (a set sequence of steps, part of a larger computer program)” and not one occurrence for each word.

Moreover, the implemented WSD procedure allows more accurate information representation. For example, let us consider the two sentences “The white cat is hunting the mouse.” and “The mouse is near the pc.” containing the “mouse” polysemous word. The disambiguation process result is showed in the following figure.

![Fig. 1: a) Disambiguation process for “The white cat is hunting the mouse.” sentence; b) Disambiguation process for “The mouse is near the pc.” sentence.](image)

After the WSD, the forum information is represented by using a list of WordNet concepts obtained through the described architecture from the forum content and their related occurrence.

For example, let us consider the following sentence to show system functionalities:
"The Semantic Web is an evolving development of the World Wide Web in which the meaning (semantics) of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content”

After the frequency reduction process we recursively evaluate for each term its synset to verify the existence of a conceptual link. If so, we have deleted an entire branch of the tree trying to get a collection of small but significant terms. The final result of the showed example consists of the single term “semantics” which accurately summarizes the initial sentence.

Considerations

Summarization can be evaluated using intrinsic or extrinsic measures; while the first methods attempt to measure summary quality using human evaluation, extrinsic methods measure the same through a task-based performance measure such as the information retrieval-oriented task (Goldstein et al., 1999). In our experiments we used an intrinsic approach analyzing W3Schools forums, official forum of the W3C (http://www.w3cforum.com/). We have performed a lot of experimental tests obtaining and elaborating a corpus of about 100 threads. This experiment is to evaluate the usefulness of concept extraction in the summarization process, by manually reading the whole thread content comparing it with automatic extracted concepts. The results clearly show that automatic concept-based summarization greatly improves the performance and produces useful information extraction supporting tutors and making learning communities effective. The extracted concepts represent a good summarization of thread contents.

Concept-based representation appears as a promising technology for implementing a distance learning environment, enabling the organization and delivery of learning materials around small pieces of semantically enriched resources (Carbonaro, 2006; Bighini et al., 2003). Items can be easily organized into customized learning courses and delivered on demand to the user, according to her/his profile and business needs (Andronico et al., 2003; Carbonaro & Ferrini, 2005).

In our experience, concept-based summarization has proven a potentially useful tool to provide a good support for tutors in virtual learning communities. To the best of our knowledge, no systems use a concept-based approach to represent online forum information in a learning environment.
REFERENCES


Rovai AP, Barnum KT. (2003), *Online course effectiveness: An analysis of student interactions and perceptions of learning*. Journal of Distance Education, 18, 57–73.


22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.


