ATRACT Workbench: An Automatic Term Recognition and Clustering of Terms[⋆]

Hideki Mima[1], Sophia Ananiadou[2], and Goran Nenadić[2]

[1] Dept. of Information Science, University of Tokyo, Japan
mima@is.s.u-tokyo.ac.jp
[2] Computer Science, University of Salford, United Kingdom
{S.Ananiadou, G.Nenadic}@salford.ac.uk

Abstract. In this paper, we introduce a web-based integrated text and knowledge mining aid system in which information extraction and intelligent information retrieval/database access are combined using term-oriented natural language tools. Our work is placed within the BioPath research project whose overall aim is to link information extraction to expressed sequence data validation. The aim of the tool is to extract automatically terms, to cluster them, and to provide efficient access to heterogeneous biological and genomic information/databases and collections of texts, all wrapped into a user friendly workbench enabling users to use a wide range of textual and non textual resources effortlessly. For the evaluation, automatic term recognition and clustering techniques were applied in a domain of molecular biology. Besides English, the same workbench has been used for term recognition and clustering in Japanese.

1 Introduction

The increasing production of electronically available texts (either on the Web or in other machine-readible forms such as digital libraries and archives) demands for appopraite computer tools that can perform information and knowledge retrieval efficiently. The size of knowledge in some domains (e.g. molecular biology, computer science) is increasing so rapidly that it is impossible for any domain expert to assimilate the new knowledge. Vast amounts of knowledge remain unexplored and this poses a major handicap to a knowledge intensive discipline.

Information retrieval (IR) either via keywords or via URL links have been used intensively to navigate through the WWW in order to locate relevant knowledge resources (KSs). While URLs can be specified in advance by the domain specialists, like links in hypertexts, IR via keywords can locate relevant URLs (and thus KSs) on the fly. URLs specified in advance are more effective in locating relevant KSs, but they cannot cope with the dynamic and evolving nature of KSs over the WWW. On the other hand, links using keywords, like in a typical IR system, can certainly cope with the dynamic nature of KSs in the WWW by computing links on the fly, but this technique often lacks the effectiveness

[⋆] This research is supported by LION BioScience, http://www.lionbioscience.com

of the direct links via URLs, as users are often forced to make tedious trials in order to choose the proper sets of keywords to obtain reasonably restricted sets of KSs. This is a well-known problem of WWW querying techniques and the techniques that combine the advantages of these two approaches are needed. Furthermore, since the URLs are often too coarse to locate relevant pieces of information, users have to go through several stages of information seeking process. After identifying the URLs of the KSs that possibly contain relevant information, they have to locate the relevant pieces of information inside the KSs by using their own navigation functions. This process is often compounded by the fact that users' retrieval requirements can only be met by combining pieces of information in separate databases (or document collections). The user has to navigate through different systems that provide their own navigation methods, and has to integrate the results by herself/himself. An ideal knowledge-mining aid system should provide a seamless transition between the separate stages of information seeking activities.

The ATRACT system, introduced in this paper, aims at this seamless navigation for the specific domain of molecular biology. It is 'term-centered', as we assume that documents are characterized by sets of technical terms which should be used as keywords for retrieval. Therefore, the very first problem to address is to recognise terms.

The paper is organised as follows: in section 2 we briefly overview ATRACT, and in section 3 we present the design of the system. In the next section we present an analysis and evaluation of our experiments conducted on corpora in the domain of nuclear receptors, and results conducted on a Japanese corpus.

2 ATRACT: an Integrated Term-Centered, Text Mining System

ATRACT (Automatic Term Recognition and Clustering of Terms) is a part of the ongoing BioPath[1] project [1]. The goal of the project is to develop software components allowing the investigation and evaluation of cell states on the genetic level according to the information available in public data sources, i.e. databases and literature. The main objective of ATRACT is to help users' knowledge mining by intelligently guiding the users through various knowledge resources and by integrating data and text mining, information extraction, information categorization and knowledge management.

As in traditional keyword based document retrieval systems, we assume that documents are characterized by sets of **terms** which can be used for retrieval. We differentiate between index terms and technical terms, and in this paper we are referring to *technical terms* i.e. the linguistic realisation of specialised concepts. In general, technical terms represent the most important concepts of a document and characterize the document semantically. We also consider contextual

[1] BioPath is a collaborative EUREKA research project coordinated by LION BioScience and ValiGen.

information between a term and its context words, since this information is important for improvement of term extraction, term disambiguation and ontology building.

A typical way of navigating through the knowledge resources on the WWW via ATRACT is that a user whose interest is expressed by a set of key terms retrieves a set of documents (e.g. from the MEDLINE database [7]). Then, by selecting the terms that appear in the document, s/he retrieves fact data from different databases in the WWW. Which databases have to be accessed should be determined automatically by the system. In order to implement the term-centered navigation described above, we have to deal with the following problems:

— **Term recognition.** In specialized fields, there is an increased amount of new terms that represent newly created concepts. Since existing term dictionaries cannot cover the needs of specialists, automatic term extraction tools are needed for efficient term discovery. In addition, naming conventions in many domains (especially in molecular biology) are highly ambiguous even for fundamental concepts (e.g. *'tumor'* can correspond to either a decease, or the mass of tissue; on the other hand, *'TsaB'* is a protein, and *'tsaB'* is a gene).

— **Selection of Databases.** There is a multitude of databases accessible over the WWW dealing with biological and genomic information. For one type of information there exist several databases with different naming conventions, organisation and scope. Accessing the relevant database for the type of information we are seeking is one of the cricial problems in molecular biology. Once the suitable database(s) is found, there is the difficulty to discover the query items within the database, as well.

ATRACT aims to provide solutions to the problems described above by integrating the following components: automatic term recognition; context-based automatic term clustering; similarity-based document retrieval, and intelligent database access.

3 ATRACT System Design

The ATRACT system contains the following components (see figure 1):

(1) Automatic Term Recognition (ATR). The ATR module recognizes terms included in HTML/XML documents using the *C/NC-value* method [2], though any method of term recognition can be used. C/NC-value method recognizes term candidates on the fly from texts which often contain unknown or new terms. This method is a hybrid approach, combining linguistic knowledge (term formation patterns) and statistics (frequency of occurrence, string length, etc). C/NC-value extracts multi-word terms and performs particularly well in recognizing nested terms i.e. sub-strings of longer terms. One of the innovative aspects of NC-value (in addition to the core C-value method) is that it is context sensitive. In a specific domain, lexical preferences of the type of context words occurring with terms is observed [6], [5]. The incorporation of contextual information is based on the assumption that lexical selection is constrained in sublanguages and that it is syntactified. The user can experiment with the results of the ATR module by tuning parameters such as threshold value, threshold rate, weights, selection of part-of-speech categories, choice of linguistic filters, number of words included in the context etc. according to his/her specific interests.

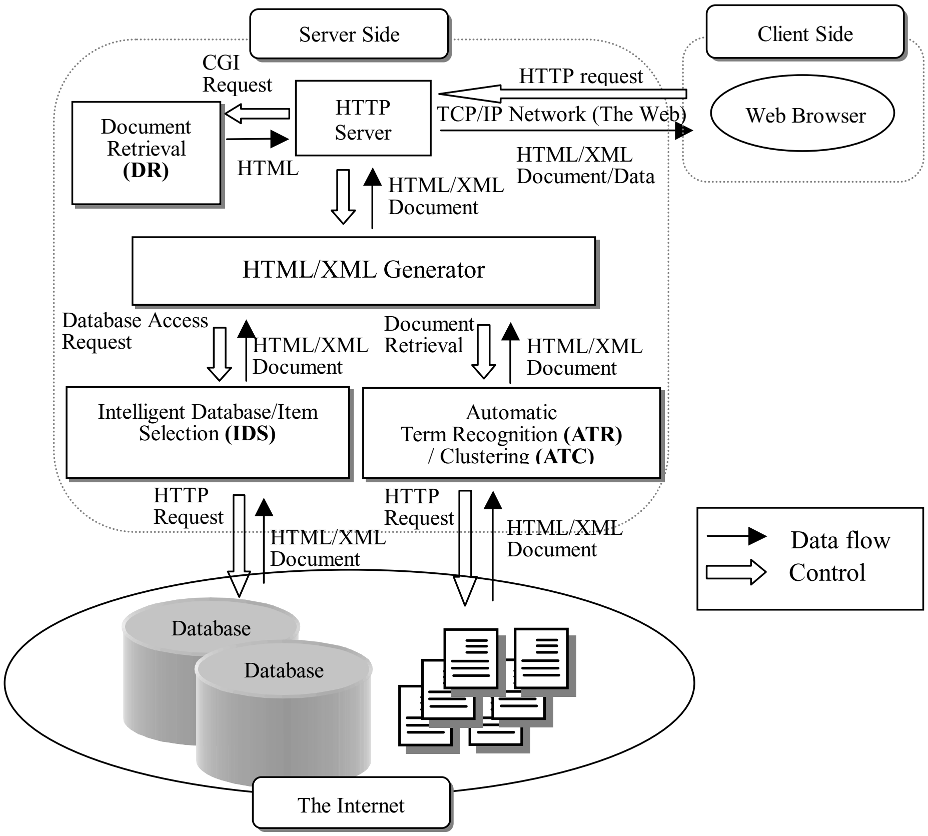

Figure 1: system design of ATRACT

(2) Automatic Term Clustering (ATC). Contextual clustering is beneficial for resolving the terminological opacity and polysemy, common in the field of molecular biology. Table 1, for example, shows problems of term ambiguity in the field. The same terms which are fairly specific and domain dependent still have several different meanings, depending on the actual context in which these terms appear. This means that, depending on the context, we have to refer to different databases to retrieve fact data of these terms.

The ATC module classifies terms recognized by the ATR module based on contextual clustering and statistical techniques. It is an indispensable component in our knowledge-mining system, since it is useful for term disambiguation, knowledge acquisition and construction of domain ontology. The approach is based on the observation that terms tend to appear in close proximity with terms belonging to the same semantic family [5]. If a context word has some contribution towards the determination of a term, there should be a significant correspondence between the meaning of that context word and the meaning of the term. Based on that observation, we compare the semantic similarities of contexts and terms. The clustering technique is based on automatically deriving a thesaurus based on the AMI (Average Mutual Information) hierarchical clustering method [12]. This method is a bottom-up clustering technique and is built on the C/NC-value measures. As input, we use bigrams of terms and their context words, and the output is a dendrogram of hierarchical term clusters.

term	protein	enzyme	compound
amino acid	+	−	−
amino acid sequence	+	−	+
pyruvate dehydrogenase	+	+	+
pyruvate carboxylase	+	+	+

Table 1: term polysemy

(3) Similarity-based Document Retrieval (DR). DR is a VSM (vector space model)-type document retrieval module. It retrieves texts associated with the current document, allowing the user to retrieve other related documents by assigning selected keywords and/or documents using similarity-based document retrieval. The user can also retrieve documents by specifying keywords.
(4) Intelligent Database/item Selection (IDS) using database (meta-) ontology. IDS selects the most relevant databases and their items using term class information assigned by ATC module and database's (meta-)ontology information. All terms are 'clickable' and dynamically 'linked' to the relevant databases over the Internet. The relevant databases should be dynamically selected according to the terms and the term hierarchy information. The module is implemented as an HTTP server, designed to choose the appropriate database(s) and to focus on the preferred items in the database(s) according to the user's requirements. The most relevant databases are determined automatically by calculating association scores between the term classes and the description of databases (such as meta-data). Furthermore, the retrieved content can be modified in order to focus on the most pertinent items for the user. It is also possible to show similar entries by calculating similarities using the term classes and the domain specific ontology when an exact matched entry is not found.

4 Experiments and Evaluation

We conducted experiments to confirm the feasibility of our proposed workbench. The evaluation was performed on 2,000 MEDLINE abstracts [7] in the domain of nuclear receptors (for English), and on a NACSIS Japanese AI-domain corpus [4]. We focused on the quality of automatic term recognition and similarity measure calculation with the use of automatically clustered terms, as all other techniques are based on term extraction.
— **Term recognition.** We have examined the performance of the NC-value method with respect to the overall performance from the viewpoint of precision

and recall by 11-point[2] score, while applying it to the same corpus and the correction set to the C-value. The top of the list produced by C-value (the first 20% of extracted candidate terms) was used for the extraction of term context words, since these show high precision on real terms. We used 30 context words for all the extracted terms in the evaluation, the number been determined empirically.

Figure 2 (left) shows the 11-point precision-recall score of NC-value method in comparison with the corresponding C-value for English. It can be observed that NC-value increases the precision compared to that of C-value on all the correspond points for recall. Similarly, NC-value increases the precision of term recognition compared to pure frequency of occurrence. Although there is a small drop in precision compared to C-value in some intervals (figure 2, right), NC-value generally increases the concentration of real terms at the top of the list.

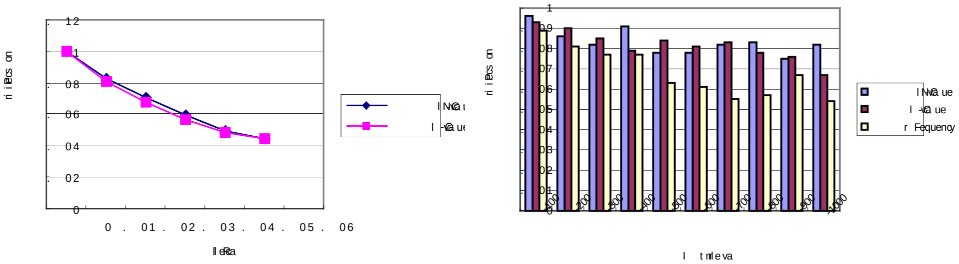

Figure 2: 11-point score (left) and interval precision (right) for English

C/NC-value method was applied on a collection of NACSIS AI-domain texts in Japanese, as well. As one can see on the figure 3, the results are similar to those obtained for English. Although the same technique is used, different linguistics filters were defined in order to describe term patterns in Japanese [9].

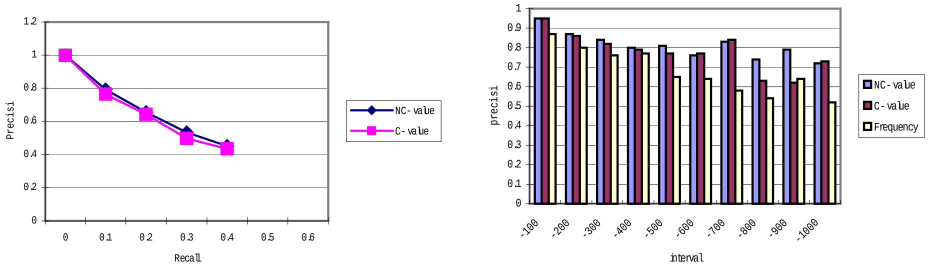

Figure 3: 11-point score (left) and interval precision (right) for Japanese

— **Clustering terms and database handling.** We used the similarity measure calculation as the central computing mechanism for choosing the most relevant database(s), determining the most preferred item(s) in the database(s),

[2] 11-point score indicates that, for example, precision at recall 0.10 is taken to be maximum of precisions at all recall points greater then 0.10.

and disambiguating term polysemy. The clustered terms were developed by using Ushioda's AMI-based hierarchical clustering program [12]. As training data, we have used 2,000 MEDLINE abstracts. Similarities between terms were calculated according to the hierarchy of the clustered terms. In this experiment, we have adopted a semantic similarity calculation method for measuring the similarity between terms described in [11]. We have used three sets (*DNA*, *PROTEIN*, *SOURCE*) of manually classified terms and calculated the average similarities (AS) of every possible combination of the term sets, that is, $AS(X,Y) = \frac{1}{n}\Sigma\ sim(x,y)$, where X and Y indicate each set of the classified terms; $sim(x,y)$ indicates similarity between terms x and y, and n indicates the number of possible combinations of terms in X and Y (except the case where $x = y$). As the table 2 shows, each AS between the same class terms, i.e. $AS(X,X)$, is greater than the others respectively. We believe that it is feasible enough to use automatic clustered terms as the main source of knowledge for calculating similarities between terms.

	DNA	*PROTEIN*	*SOURCE*	# of terms
DNA	**0.533**	–	–	193
PROTEIN	0.254	**0.487**	–	235
SOURCE	0.265	0.251	**0.308**	575

Table 2: average similarities

However, despite these results on clustering and disambiguation, searching through suitable databases on the Web still remains a difficult task. One of the main problems encountered is that we are not certain which databases have items which best describe our request, i.e. we do not know whether the related databases are pertinent to our request. In addition, the required information is sometimes distributed into several databases, i.e. almost all databases are disparate in terms of information contained.

5 Conclusion and Further Research

In this paper, we have presented ATRACT, a web-based integrated text and knowledge mining workbench. ATRACT extracts automatically terms based on a combination of linguistic and statistical knowledge, clusters terms and provides seamless navigation and access to heterogeneous databases and collections of texts. The workbench provides a user with a friendly environment for term extraction and clustering from a variety of knowledge and textual sources. The system enables a logical integration of databases on the Web: the design allows the users to refer to the required items gathered from several web databases as if they access certain sophisticated single database virtually.

Important areas of future research will involve improvement of term recognition using semantic/clustered term information and additional syntactical structures (e.g. term variants and coordination [3], [10]), and improvement of database handling. Since our goal is dealing with the 'open world' of databases, due to insufficiency of information on what sort of data is contained in each database, selecting the most associative databases is one of the crucial problems. Therefore we will have to resolve problems of choosing database(s) from large amounts of databases on the Web (and to recognise newly launched databases as well), and to modify the 'view' of each database according to the requirements (since the format styles vary from site to site). We expect that meta-data information could be useful to select database(s) if enough meta-data about each database is available on the Web. Regarding the format style of each database, we expect that the popularity of XML might be a solution of the problem.

References

1. Ananiadou, S., Albert, S., Schuhmann, D.: "Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline", Genome Informatics Series, vol.11, 2000
2. Frantzi, K. T., Ananiadou, S., Mima, H.: "Automatic Recognition of Multi-Word Terms: the C-value/NC-value method", International Journal on Digital Libraries Vol. 3, No. 2, pp.115–130, 2000
3. Kehagia, K., Ananiadou S.: "Term Variation as an Integrated Part of Automatic Term Extraction", in Proc. of 22nd Conference for Greek Language, Thessaloniki, Greece, 2001 (forthcoming)
4. Koyama, T., Yoshiokka, M., Kageura, K.: "The Construction of a Lexically Motivated Corpus — the Problem with Defining Lexical Units", in Proc. of LREC 1998, pp.1015-1019, Granada, Spain, 1998
5. Maynard, D., Ananiadou, S.: "Identifying Terms by Their Family and Friends", in Proc. of 18th International Conference on Computational Linguistics, COLING 2000, pp.530–536, Luxembourg, 2000
6. Maynard, D., Ananiadou, S.: "TRUCKS: a Model for Automatic Term Recognition", in Journal of Natural Language Processing, Vol. 8, No. 1, pp.101–125, 2001
7. MEDLINE, National Library of Medicine, `http://www.ncbi.nlm.nih.gov/PubMed/`
8. Mima, H., Ananiadou, S., Tsujii, J.: "A Web-based Integrated Knowledge Mining Aid System Using Term-oriented Natural Language Processing", in Proc. of The 5th Natural Language Processing Pacific Rim Symposium, NLPRS'99, 13-18, 1999
9. Mima, H., Ananiadou, S.: "An Application and Evaluation of the C/NC-value Approach for the Automatic Term Recognition of Multi-Word in Japanese", in Terminology 6:2, 2001 (forthcoming)
10. Nenadic, G.: "Local Grammars and Parsing Coordinaton of Nouns in Serbo-Croatian", in Text, Speech and Dialogue - TSD 2000, Lecture Notes in Artificial Intelligence 1902, Springer Verlag, 2000
11. Oi K., Sumita E., Iida H.: "Document Retrieval Method Using Semantic Similarity and Word Sense Disambiguation in Japanese", Journal of Natural Language Processing, Vol.4, No.3, pp.51-70, 1997
12. Ushioda A.: "Hierarchical Clustering of Words", In Proc. of COLING '96, Copenhagen, Denmark, 1996