Immune-Based Framework for Exploratory Bio-information Retrieval from the Semantic Web

Doheon Lee[1,*], Jungja Kim[2], Mina Jeong[2], Yonggwan Won[2], Seon Hee Park[3], and Kwang-Hyung Lee[1]

[1] Dept. of BioSystems/AITrc, KAIST, Daejeon, Korea
[2] Dept. of Computer Engineering, Chonnam Nat'l Univ., Gwangju, Korea
[3] Computer and Software Research Lab., ETRI, Daejeon, Korea

Abstract. This paper proposes an immune-based framework for adaptive query expansion in the semantic web, where exploratory queries, common in biological information retrieval, can be answered more effectively. The proposed technique has a metaphor with negative selection and clonal expansion in immune systems. This work is differentiated from the previous query expansion techniques by its data-driven adaptation. It utilizes target databases as well as the ontology to expand queries. This data-driven adaptive feature is especially important in exploratory information retrieval where the querying intention itself can be dynamically changed by relevance feedback.

1 Introduction

The semantic web is a collaborative effort to make the web information machine-understandable. By associating semantic descriptions with web information elements, computer programs such as web robots can access the web in more meaningful ways. For example, suppose that an XML (eXtensible Markup Language) document has 'Title' and 'Author' elements. Without extra information, the XML document alone can indicate at best those two elements are components of a certain information unit [1]. If we associate an RDF (Resource Description Framework) description such that 'Author Publishes Title' with the XML document, a web robot can infer that those two elements are semantically related to each other by 'Publishes' relationship. If we associate a different description such that 'Author Has Title' with the same document, it leads different interpretation. Since this sort of semantic treatment is an essential ingredient to leverage the web as an information thesaurus, there have been fast growing interest and efforts to develop semantic web frameworks and RDF-based ontology systems recently [2][3][4].

Currently, we have over 500 bio-databases available in the Internet [5], and tons of bio-information repositories containing experimental summaries, published papers, and white papers. It is quite common that information requests arising in biology laboratories can be answered by integrating different but related information from such disparate sources. Thus, there have been many research and development endeavors to achieve integrated information retrieval systems, whichever they are virtual integration or data warehousing-based [6][7]. In the database community, the integration of heterogeneous data sources has been a longstanding but still a difficult

* To whom correspondence should be addressed

J. Timmis et al. (Eds.): ICARIS 2003, LNCS 2787, pp. 128–135, 2003.

issue. One of the most practical approaches is to pursue standardization as far as possible both in syntax and semantics so that information systems understand and exchange different information effectively. Despite that so many proposals have come and gone, XML and RDF-based semantic web is the most promising solution for that purpose now. XML facilitates to organize the information in a syntactically standardized way, and RDF makes it possible to fill the gaps between syntax and semantics.

Though information retrieval from bio-information repositories shares many common characteristics with traditional information retrieval domains such as library index systems and web search engines, it reveals at least two significant differences. Firstly, it is much harder to formulate the user query precisely at once since complex domain knowledge is heavily involved [10]. Rather, most users rely on interactive query reformulation based on relevance feedback [11]. Even if the users are well-trained to select proper terms to describe their queries, the severe diversity of biological terms from various disciplines causes significant amount of false positives and false negatives. More noticeable point is that the querying intention itself can be dynamically changed by relevance feedback. The second difference alleviates the problem. There are a lot of active standardization efforts on biological ontology construction [7][8][9], and more and more autonomous bio-information repositories have begun to accept the necessity and usefulness of such ontologies.

To remedy the hardness of precise query formulation, query expansion or query generalization techniques have been studied actively especially in the Internet-based information retrieval [12][13][14][15][16]. Though there have been many proposals for query expansion, they have common trade-offs. If queries are expanded too broadly, they retrieve too many false positives, and suffer unnecessary access cost. If queries are expanded too narrowly, they end up with many false negatives. Thus, the key issue of query expansion is how to expand queries to focus only on relevant information. To challenge this issue, it is necessary to expand queries in an adaptive manner, which means queries are expanded based on the result feedback. Recently, query expansion techniques for bio-databases focusing on vagueness of query expressions have been proposed [10][17]. However, they are not adaptive in the sense that they do not utilize relevance feedback. We have found a procedure of adaptive query expansion can have strong analogy with negative selection and clonal expansion principle in natural immune systems. Based on this observation, this paper proposes an adaptive query expansion technique for explorative bio-information retrieval from the semantic web in the framework of artificial immune systems. This approach can be regarded to belong to the information immune systems, which utilize the immune metaphor to filter unnecessary information glut [18].

Section 2 explains the metaphor between the proposed technique and immune principles, and Section 3 presents the proposed technique along with illustrative examples. Section 4 proposes a semantic web-based system architecture for implementing the proposed technique, and Section 5 concludes.

2 Immune Metaphor for Query Expansion

Let us remind the procedure of natural immune systems briefly to provide common terminology in this paper [19][20]. We will use the term ‘antibodies’ instead of ‘B cells’, and omit unnecessary biological details to focus on fundamental principles. (i)

In a primary lymphoid such as bone marrows, a large variety of immature antibodies are produced by combinatorial recombination of gene segments; (ii) Those antibodies who can bind with self-antigens are eliminated, i.e. negative selection; (iii) Mature antibodies circulate the body, and bind with foreign antigens; (iv) The foreign antigen-bound antibodies proliferate, and undergo hypermutation to produce more effective antibodies, i.e. clonal expansion; (v) Some of those antibodies differentiate to memory cells to cope with future infection, while the others become effector cells that can immediately bind with another instances of foreign antigens.

Our proposed information retrieval procedure follows similar steps. (i) Given a user query q_0, the procedure produces a set of queries Q_0 by expanding q_0 based on given ontologies; (ii) If the user-side querying system maintains a set of rules to avoid unnecessary information, the matching queries in Q_0 will be removed; (iii) The resulting Q_0 will be distributed to local information repositories through the Internet; Some of queries in Q_0 matches with data units in each local repository; (iv) Those matched queries are reported to the user-side querying system; Then subsequent query expansion is applied each query set from each local repository based on the given ontologies; (v) The matched queries are integrated into the knowledge-base to facilitate effective query expansion in another query processing. Table 1 shows an immune metaphor for our information retrieval technique.

Table 1. Immune metaphor for the proposed adaptive query expansion

	Natural Immune Systems	Adaptive Query Expansion
(i)	Gene recombination	Initial query expansion
(ii)	Elimination of improper antibodies	Elimination of uninteresting queries
(iii)	Binding to foreign antigens	Matching with data in local repositories
(iv)	Proliferation	Subsequent query expansion
(v)	Memory cells	Storing useful queries

3 Adaptive Query Expansion Procedure

To apply the adaptive query expansion, we suppose that local bio-information repositories provide RDF/XML signatures representing their respective information. Table 2 shows an example.

Table 2. Part of database signatures from local information repositories

No.	Substance	Relation	Process
d_1	Metabolic Enzyme	Related-to	Stomach Cancer
d_2	p53	Related-to	Cell cycle
d_3	E. coli	Exhibit	Chemotaxis
d_4	GPCR	Mediate	Signal Transduction
d_5	Metabolic Enzyme	Activate	Stomach Cancer

Though actual information repositories contain much more signatures than depicted in Table 2, it is given just for illustration. Actual representation of the signatures is supposed to be in RDF/XML. Figure 1 shows the appearance of such representation for the first and second rows of Table 2 as examples.

```
<rdf:RDF>
   <info:unit rdf:about="http://mydata.org/dbsig#d1">
        <info:substance> Metabolic Enzyme </info:substance>
        <info:relation> Related-to </info:relation>
        <info:process> Stomach Cancer </info:process>
   </info:unit>
   <info:unit rdf:about="http://mydata.org/dbsig#d2">
        <info:substance> p53 </info:substance>
        <info:relation> Related-to </info:relation>
        <info:process> Cell Cycle </info:process>
   </info:unit>
</rdf:RDF>
```

Fig. 1. RDF-XML representation of signatures in local information repositories

Also suppose that we have ontology as in Figure 2. Arrows and tildes represent 'isa' and 'synonym-of' relationships respectively.

```
Bio-molecule => Protein | Nucleic Acid | Compound
Protein => Cytoplasmic Protein | Nucleus Protein
Cytoplasmic Protein => Metabolic Enzyme | Protein Synthesis
    Factor
Disease => Infection Disease | Cancer | …
Cancer => Prostate Cancer | Stomach Cancer | Lung Cancer | …
Stomach Cancer ~ Gastric Cancer
Related-to => Activate | Repress
```

Fig. 2. Ontology example for adaptive query expansion

There are a lot of ongoing efforts to construct ontologies for various biological domains including genomics, proteomics, and medicine. Figure 3 shows a part of the ontology in Figure 2 in the form of RDF/XML as in the Gene Ontology [9].

Now, let us suppose that we have an initial query q_0 such as "Find information about cytoplasmic proteins related to stomach cancer." It is a typical query example in biology laboratories where stomach cancer is studied in the molecular level. Such queries have exploratory intention in nature. Though the query indicates cytoplasmic proteins and stomach cancer specifically, neighbor information such about intestinal cancer or membrane proteins could be also useful to the query submitters. It would become much more important if there are significant interactions between cytoplasmic proteins and membrane proteins in stomach cancer development.

Thus, query expansion to encompass such neighbor information is quite useful in exploratory query processing. By referring to the given ontology, the initial query q_0 can be expanded to a query set Q_0 as in Figure 4.

This query expansion is done by substituting terms in the initial query with direct descendent or direct ancestor terms in the given ontology one by one. As this illustrative example contains a few data and unrealistically small ontology, the expansion might look trivial work. However, it can come up with much more expanded queries in actual situations where the amount of data is huge and the size of ontology is realistic. For example, the current gene ontology in [9] contains over 70,000 terms. It can

be mapped into gene recombination to produce a diversity of antibodies in the framework of the immune metaphor. As V/J/D regions of the antibodies are selected from gene segment libraries, this query expansion is done using ontology libraries.

```
<rdf:RDF>
   <go:term rdf:about="http://www.ontology.org/go#0001">
      <go:accession>GO:0001</go:accession>
      <go:name>Bio-Molecule</go:name>
   </go:term>
   <go:term rdf:about="http://www.ontology.org/go#0002">
      <go:accession>GO:0002</go:accession>
      <go:name>Protein</go:name>
      <go:isa rdf:resource="http://www.ontology.org/go#0001">
   </go:term>
   <go:term rdf:about="http://www.ontology.org/go#0234">
      <go:accession>GO:0234</go:accession>
      <go:name>Stomach Cancer</go:name>
      <go:synonym>Gastric Cancer</go:synonym>
   </go:term>
</rdf:RDF>
```

Fig. 3. RDF/XML representation of ontology

- q_{01}: Cytoplasmic Proteins, Related-to, Stomach Cancer (=q_0)
- q_{02}: Metabolic Enzyme, Related-to, Stomach Cancer
- q_{03}: Protein Synthesis Factor, Related-to, Stomach Cancer
- q_{04}: Metabolic Enzyme, Repress, Stomach Cancer

Fig. 4. Q_0: Expanded queries from the initial query q_0

Let us further suppose that the user is not interested in metabolic enzymes repressing stomach cancer development. Then, the user can specify this constraint as his/her own preference information. According to the specified user constraint, one of the expanded queries q_{04} is eliminated from the query set before distributed to local information repositories. This pre-screening can be mapped into negative selection to remove such antibodies that bind with self-antigens.

Among the expanded queries in Figure 4, only q_{02} is matched with data d_1 in Table 2. In the immune metaphor, this can be regarded as an antibody binds with a foreign antigen. The matched query q_{02} is further expanded by referring to the ontology as in Figure 5.

- q_{11}: Metabolic Enzyme, Activate, Stomach Cancer
- q_{12}: Metabolic Enzyme, Repress, Stomach Cancer

Fig. 5. Q_1: Expanded queries from the matched query, q_{02}

Again, q_{11} is matched with data d_5 in Table 3. After keeping the matched result, q_{11} is further expanded again by referring to the ontology as in Figure 6.

- q_{21}: Metabolic Enzyme, Activate, Gastric Cancer

Fig. 6. Q_2: Expanded queries from the matched query, q_{11}

This iterative process will continue until the user is satisfied with the result up to the point, or no more additional information can be retrieved.

4 Semantic Web-Based System Architecture

The proposed immune-based information retrieval system is aiming at web-based distributed information systems where each local information unit is characterizing its own information content in terms of RDF/XML signatures. Figure 7 depicts a system architecture enabling the proposed information retrieval procedure.

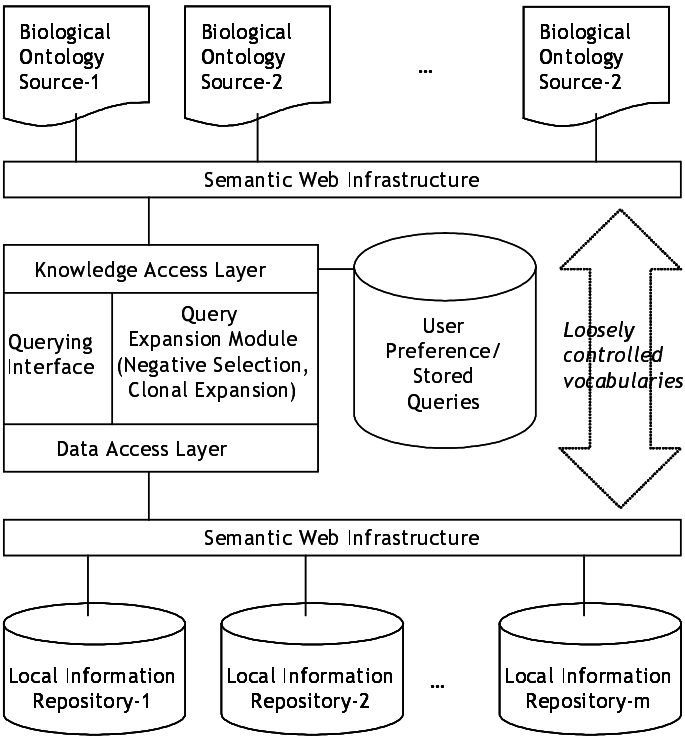

Fig. 7. Semantic web-based system architecture for adaptive query expansion

There are a variety of biological ontology sources available through Internet, which includes Molecular Biology Ontology (MBO), Gene Ontology (GO), and TAMBIS Ontology (TaO) [8]. Since most of them provide RDF-based representation to facilitate machine processing, they are already prepared to join the semantic web commu-

nity. In addition, most of major bio-databases already support XML-based representation, and are on the way to support RDF.

As explained in Section 3, Query Expansion Module has negative selection sub-module to pre-screen unnecessary queries after each expansion. For such negative selection, semantic rules describing unnecessary information should be stored in the User Preference database. It also has clonal expansion sub-module to expand useful queries. Instead of maintaining local ontology for query expansion, we propose to utilize online ontology sources since the ontology itself is growing fast. Furthermore, we can expect that data elements in local information repositories are controlled loosely by major ontology sources since those information repositories begin to refer to major ontology sources. This loose relationship between local information repositories and major ontology sources play an important role for improving the information retrieval effectiveness. Selected queries from each expansion step can be stored in the Stored Queries database to be used directly in the next query processing. This feature can be mapped into memory cells in the immune metaphor.

5 Concluding Remarks

This paper has proposed an immune-based framework for adaptive query expansion in the semantic web, where exploratory queries, common in biological information retrieval, can be answered more effectively. The proposed technique has a metaphor with negative selection and clonal expansion in immune systems. This work is differentiated from the previous query expansion techniques by its data-driven adaptation. It utilizes target databases as well as the ontology to expand queries. This data-driven adaptive feature is especially important in exploratory information retrieval where the intention of querying itself can be changed by result feedback.

We are going to implement the proposed system using mobile agent technology. Since the information retrieval procedure has inherent parallelism, it fit quite properly into mobile agents. Mobile agents equipped with standard communication capability with ontology sources can navigate through the semantic web, proliferate upon their own matching situations, and extract relevant information to the users.

References

1. K. Sall, XML Family of Specifications: A Practical Guide, Addison-Wesley, 2002
2. T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific American, May 2001
3. C. Bussler, D. Fensel, and A. Maedche, “A Conceptual Architecture for Semantic Web Enabled Web Services,” SIGMOD Record, 31(4), pp. 24-29, 2002
4. S. Decker, et al, “The Semantic Web: the Roles of XML and RDF,” IEEE Internet Computing, 4(5), pp. 63-73, 2000
5. C. Discala, et al, “DBcat: a Catalog of 500 Biological Databases,” Nucleic Acids Research, 28, pp. 8-9, 2000
6. D. Buttler, et al, “Querying Multiple Bioinformatics Data Sources: Can Semantic Web Research Help?” SIGMOD Record, 31(4), pp. 59-64, 2002

7. C. Goble, et al, "Transparent Access to Multiple Bioinformatics Information Sources," IBM Systems Journal, 40(2), pp. 532-551, 2001
8. I. Yeh, P. Karp, N. Noy, and R. Altman, "Knowledge Acquisition, Consistency Checking, and Concurrency Control for Gene Ontology (GO)," Bioinformatics, 19(2), pp. 241-248, 2003
9. The Gene Ontology Consortium, "Gene Ontology: Tool for the Unification of Biology," Nature Genetics, 25, pp. 398-416, 2000
10. Y. Chen, D. Che, and K. Aberer, "On the Efficient Evaluation of Relaxed Queries in Biological Databases," Proc. of ACM Conference on Information and Knowledge Management, McLean, 2002, pp. 227-236
11. G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," J. of American Society for Information Science, 41, pp. 288-297, 1990
12. H. Joho, et al, "Hierarchical Presentation of Expansion Terms," Proc. of ACM Symposium on Applied Computing, Madrid, 2002, pp. 645-649
13. Y. Kanza, W. Nutt, and Y. Sagiv. "Queries with Incomplete Answers over Semistructured Data", Proc. of ACM PODS, Philadelphia, May, 1999.
14. T. Gaasterland. "Cooperative Answering through Controlled Query Relaxation", IEEE Expert, 12(5), pp. 48–59, 1997
15. M. Mitra, A. Singhal, and C. Buckley. "Improving Automatic Query Expansion," Proc. of ACM SIGIR, Melbourne, Aug. 1998
16. A. Motro. "Query Generalization: A Method for Interpreting Null Answers", Proc. of Expert Database Systems Workshop, Kiawah Island, Oct. 1984
17. D. Che, Y. Chen, K. Aberer, "A Query System in a Biological Database," Proc. of 11th International Conference on Scientific and Statistical Database Management, Cleveland, 1999, pp. 158-168
18. D. Chao and S. Forrest, "Information Immune Systems," Proc. of the 1st Int'l Conf. on Artificial Immune Systems, 2002
19. L. de Castro and J. Timmis, Artificial Immune Systems: A New Computational Intelligence Approach, Springer, 2002
20. D. Dasgupta, Artificial Immune Systems and Their Applications, Springer, 1998