Text Mining in Bioinformatics by Software Agents

Hadee Faiiazee[1], S.A.R. Al-Haddad (Dr.)[2], Rusli Abdullah (Associate Prof. Dr.)[3], Khairulmizam Samsudin (Dr.)[4]

[1]PhD Candidate, Department of Computer and Communication System Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Selangor, Malaysia; E-mail: h.faiiazee@gmail.com
[2] Department of Computer and Communication System Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Selangor, Malaysia;E-mail: sar@eng.upm.edu.
[3]Department of Computer and Communication System Engineering, Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang, Selangor, Malaysia; E-mail: rusli@fsktm.upm.edu.my
[4]Department of Computer and Communication System Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Selangor, Malaysia; E-mail: kmbs@eng.upm.edu.my

Abstract

New high throughput methods resulted in huge data amount in biology. The main part of related information for genomic data is the peer-reviewed papers. Text is accessible in bulky quantities, but is often messy and disorganized. The only practical answer to analyze these data sets is using computational methods that are unrestricted by the quantity of data. Hence, using text mining tools and innovative knowledge mining software that supports the curators is inevitable. Software agents are one of these kinds of tools that provide new solution for text mining. This new solution has its individual advantages in bioinformatics. Their ability to being mobile, highly customize ability and interaction aspect of agents, are some of these advantages. In this paper we offer a software agents frame for data mining in bioinformatics. Evaluation of agents performance on text mining duty is one of significant component of our workflow.

1. Introduction

Autonomous action and interaction capability are main features of software agents. Software agents as a new area in computer science, integrate with other technologies to promote technologies to higher levels. One of these types of integration is data mining driven agents. Some difficult issues in each side can be successfully addressed through this interaction.
Data mining needs software agents flexibilities to deal with distributed, dynamic, unpredictable and open environments data mining. On the other hand, adding data mining capabilities to agents makes them smarter and more adaptable that results improved performance in agents [1].
Bioinformatics brings computational tools in biology context. Since most important part of related information for genomic data is textual data [2], data mining is one of those tools that have been used to mine, integrate and extract knowledge from exponential growing amount of textual data in this field.
In this paper we have proposed a framework for textual data mining by software agents in bioinformatics. In the next section, we review software agents in bioinformatics field. Then after we present our methodology on how to use data mining driven agents on bioinformatics text data.

2. Background

This section of paper is not a complete review on software agents in bioinformatics, but we tried to highlight important phases of software agents emergence in bioinformatics.
One of the best fields for agent integration is biology and one of the best parts of biology that agents can take participation is information process automation [3].
Early on 2000 decade, software agents appeared first in bioinformatics. GeneWeaver was the first project that used agent idea in bioinformatics [4]. Later, on 2001, using software agents as data mining tools started on DNA-microarray data integration by angeletti, culmone and merelli [5]. At the same time and regarding that just a small number of multi agents systems planned for bioinformatics, first Multi agent systems appeared in bioinformatics by the name of SPT for automated genomic annotation by decker, zheng and schmidt [6]. In 2002 one of the most important aspects of software agents, being mobile, came to biology by proposing Bioagent [7].
Using software agents as independent bioinformatics tool started about 2002, for example in [8] and [9]. In 2005, the potential advantages of integrating bioinformatics and multi-agent systems (MAS) were recognized by bazzan and others [10]. BioWMS [11] was the first agent with web-based interface that launched in 2007. At the same year and in very specific field, text mining agents appeared in bioinformatics [12].

3. Text Mining Agents in Bioinformatics

Around 2000, the concept of data mining came to biology. Since then lots of data mining algorithms, tools, softwares and web tools on bioinformatics have been offered. Software agents, as a specific tool, have been considered for data mining from 2001.
In biology, the most reliable resource of information and data format is textual scientific literature. The highest importance part of knowledge in this field is the peer-reviewed published papers [13]. Without new developed bioinformatics tools and methods, acquitting with new intensive-throughput techniques and advanced technologies that provide a massive sum of articles, basically is out of reach. These new tools should be able to integrate different data from different sources, like experimental data and biological databases, aiming to speed up discovery of knowledge and curation of databases [14].
Until now, deployment of text mining full capabilities in biology is remained out of reach. Achieving goals like mining from heterogeneous data sources, full article text mining and multidimensional complex mining, face text data mining in bioinformatics with new challenges.
Software agents are in capable of addressing some issues that bioinformatics deals with them in text data mining. In the following, we reviewed some of them:

A. Integration, Communication and Collaboration

In biology, the required data usually is located among distributed sources in a dynamic environment that provide very heterogeneous materials [15]. Although experts tried to find solutions to this with centralization and decentralization approaches, so far no solution with satisfactory results presented [16].
Agents are able to integrate multiple data bases in distributed locations with different data formats [17]. In bioinformatics and in lack of such tools, since researchers repeatedly have to deal with various databases and web servers in parallel, it a scary task to integrate non-homogeneous data from dissimilar and remote databases linked with many web servers.

B. Usability and Interactivity

In end user point of view, usability is one of most important aspect of new tools and techniques. It will be more important in the case that we have lots of users with different backgrounds like pharmacology, biomedicine, bioinformatics, biology, and with little or no knowledge about text mining and natural language processing techniques.
A different group of users, the 'content providers' including curators, has their own definition of interaction, focusing more on drilling down for proofs, association between resources, and generating new data resources to get new perceptions [18].
While we want high-level interactive tools and current text mining tools in bioinformatics are not efficient at interactive systems that users could adopt easily [18], agents are accepted having successful capabilities of intelligent interactive dealing with different users [19].

C. Dynamic Customization

Different classes of users, use bioinformatics tools for different means. Therefore successful tools are those ones that could cover different individual needs specifically.
Customization in agents means they could employ a lot of logic, queries and abbreviation techniques depending on the condition and users requirements [20].
Some experts argue that the next generation of text-mining applications should be more user-focused [16][21]. Interactive customization character of agents represents them as tools with unlimited flexibility that could be a respond to this requirement.

4. Proposed methodology

In this part, we propose a framework for having software agents advantages in biology textual mining. The proposed methodology is consisting of a workflow that starts with designing agents for data mining function, creation relevance ontology, agent functionality development and agent conception based on knowledge models for agents deferent categories. Incorporate the knowledge forms to agents, multi-agent application building and watching agent actions are the next phase. Assessment of agents performance is the last phase (Fig 1).

A. Designing Agents

Designing agents are including defining their main features. Degree of autonomy and communication ability is two important of them. In our methodology ability to trade information and interactivity are important too. Other main features that should be respected are mobility for capable of integration between distributed sources, ability of learning, and cooperativeness. Cooperativeness in text data mining brings more intelligence to swarm of agents to carry out complicated and advanced mining like full paper mining that is one of the current text mining challenges in bioinformatics.

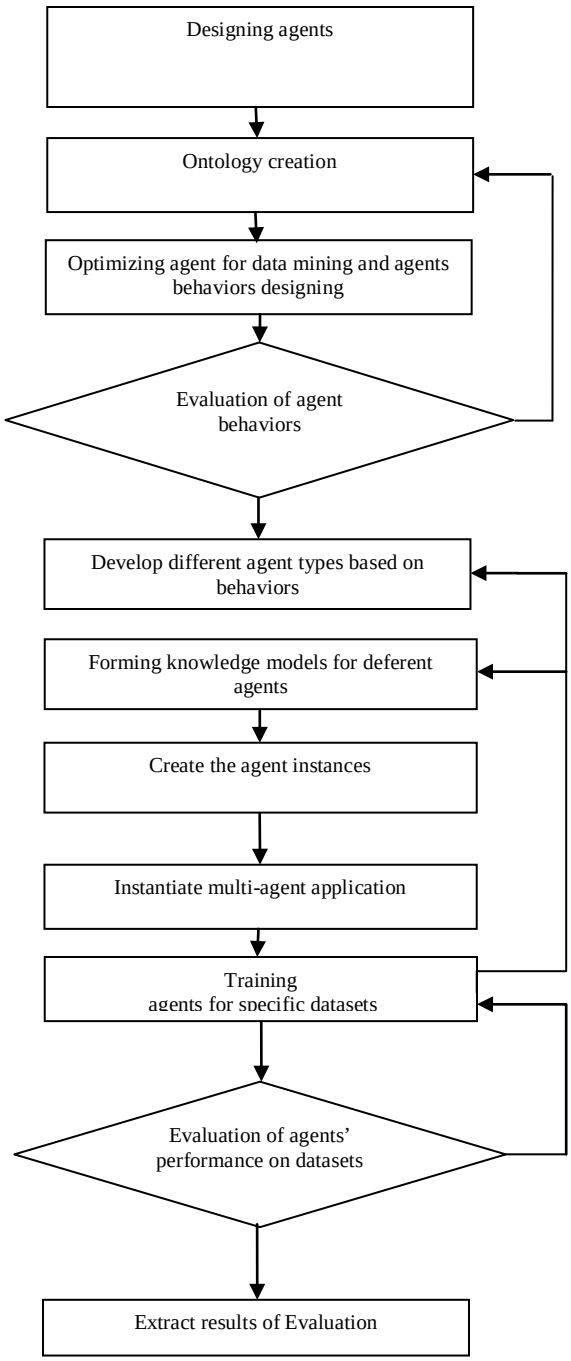

Fig. 1 Methodology workflow

B. Ontology Creation

Ontology in biology symbolizes an important source of representative knowledge. Ontology in agents creation will be used as conjunction between domain knowledge and data mining techniques. In proposed methodology, ontology are including biomedical entities like gene and protein, their names, like basal ganglia, relevant knowledge, like "cystic fi brosis is caused by a mutation of the CFTR gene located on chromosome 7" . In bioinformatics, most recent researches use ontology for knowledge prior base that get it easier to some data mining techniques like clustering [22].

C. *Optimizing Behaviors Designing*

One outcome of ontology defining is using ontology to establish set of regulations, which will be used to create the system output in the form of intended. It will assist agent development to be more optimized according to certain domain, in this case biology text mining. The regulations will apply by define and/or optimize agents behaviors.

D. *Behaviors Evaluation*

Agent behaviors after designing and/or optimizing should be evaluated according to system anthology. One result of evaluation could be ontology revising and/or behaviors modifying. The evaluation will go on until getting satisfying result of agent performance.

E. *Different Agent Developing*

Deferent behaviors in general means we need to different agents. It owes to have conflict between behaviors in taking priority or in doing parallel jobs. In addition, in some occupations, we need specific agents with different tasks and behaviors.

F. *Forming Knowledge Models*

Dealing with deferent databases, that usually means different data formats, is a challenge in bioinformatics, as mentioned before. In addition, flexibility in agents interaction means agents face with different user queries and dynamic environment. In all these cases we need different form of ontology, or as we name it in this phase "knowledge model". Knowledge model is simplified instances of main ontology model to apply to different agents. Dynamic knowledge model is the most advanced model of knowledge that gives the highest rate of usability and performance.

In bioinformatics environment, most of data bases dictate their own rules on entities naming like gene naming, data saving and data representing. Some of them present data in specific format and in specific relation rules with other parts of datasets or other data bases. Adding meta data to main data is one example of this. Lots of abbreviations and acronyms is another instance. Addition to this, some data mining algorithms show better mining results on specific datasets than the others [23].

All of these matters should be considered in this section.

G. *Create the Agent Instances*

Concepts shaped in previous phases come to action in this step by creating different agents instances with different types and behaviors. Selecting appropriate framework for agent development is one of the important issues in this phase.

Need to have more than one agents at the same time, bring the methodology to point that need to multi agents platform is inevitable. Working with different database with heterogeneous datasets in distributed environment, having parallel task to do with together in the same time and providing services for more than one client, are some kinds of need to have multi agents system.

Agents interaction with each other makes the multi agent system more efficient and intelligent; but bring some complexity to design optimized cooperative environment.

Communication specification and interaction protocols are two important issues that should be taken into account in this case.

H. *Training*

Training created agents is the last step before putting agents in real action. Training phase provide opportunity to last evaluation of agents, and also their knowledge model. This phase could provide some feedback, if necessary, to agents design as well as knowledge model phases.

Training phase also could be recurring step for agents after starting exploiting, in the case of utilizing agents for client in different sequential queries.

I. *Evaluation*

Evaluation is the most important phase of proposed methodology. Precision and recall are the most commonly used measures for evaluating text mining systems.

Nevertheless in the case of bioinformatics text mining, an evaluation criterion defining and acceptable data formats are critical to assess performance of methods and algorithms. To satisfy this issue, we are witness of raising a number of challenges on text mining via task-based procedures.

Furthermore In the case of agents, a robust base establishment for evaluating the efficiency of agents and multi-agent systems, is one of the important remaining open issues in this perspective [24].

However, evaluation of agents performance on tree level, between agents and current successful approaches in international challenges on text mining in bioinformatics, between different data mining techniques and between different data sets could be a good point to start.

5. Discussion

A wide collection of tools for text mining are presented today. These Tools propose optimized or specific queries on broad range of data. Every existing tool has special features, and may be ideal for specific users or to deal with particular problems.
In proposed methodology some issues should be considered. One concern is choosing agent developing framework. Some free and open source frameworks are available for design and developing agents (for instance Agent academy: http://sourceforge.net/projects/agentacademy).
The other concern is on software agents evaluation. To this time, there is no well qualified method that enables mining of related information from the biology textual data by automated algorithms [25]. Defining corresponding assessment measures and common data formats is critical to verify the presentation of different technique and methods applied in biomedical text mining. On the other hand, one of the essential mutual issues in agent-mining is assessment issues such as technical implication. Therefore, evaluation is one of the open issues in this methodology. Evaluating text mining via task-based challenges could be an answer that in recent years has gotten a lot of interest among experts.
However and according our information, the system suggested here offers, at least, three special features that, are not addressed in any other tool.
First, software agents are computational units that provide great flexibility on their tasks, so we could have significance of dynamic knowledge models on our text mining. This feature is more significant on using mining with special concerns on singular data sets. The ability of users to have different behaviors on agents, permits users to investigate on literature from particular views. Different weights on different keywords are one of this feature results.
Second, mobility make possible to automatically develop the primary task on physically remote databases that is one of the bioinformatics data sets character.
Third, since agents assessment is still an unsolved concern, our system uses a valuation technique that could be used to agents assessment.

6. Conclusion

In this paper we reviewed benefits of data mining by software agents on text datasets in bioinformatics. We offer a workflow to build up agents for data mining for bioinformatics. Evaluation of agents performance on text mining duty is one of significant component of our workflow.
Additional development will be on the evaluation of agents performance, as an remaining unsolved concern in agents world. We considered carrying out further evaluation and comparison of this methodology with other conventional methods.

7. References

1. Zhang, C., Z. Zhang, et al. (2005). "Agents and data mining: mutual enhancement by integration." Autonomous Intelligent Systems: Agents and Data Mining: 259-275.
2. Krallinger, M., F. Leitner, et al. (2010). "Analysis of biological processes and diseases using text mining approaches." Methods in Molecular Biology 593: 341-382.
3. Luck M, McBurney P, Preist C. Agent technology: enabling next generation computing (a roadmap for agent based computing). Technical report, AgentLink II, 2003.
4. Bryson K, Luck M, Joy M, Jones D. Applying agents to bioinformatics in Geneweaver. In: Cooperative Information Agents IV, Lecture Notes in Artificial Intelligence. Berlin/ Heidelberg: Springer-Verlag, 2000: pp. 60–71.
5. Angeletti M, Culmone R, Merelli E. An intelligent agent architecture for DNA-microarray data integration. In NETTAB ç CORBA and XML:Towards a bioinformatics integrated network environment, Genova, Italy, 2001.
6. Decker, K., Zheng, X. and Schmidt C.J. SPT: A Multi-Agent System for Automated Genomic Annotation (2001) Automated Agents. 5, 433-440.
7. Merelli E, Culmone R, Mariani L. Bioagent: A mobile agent system for bioscientists. In: NETTAB ç Agents in Bioinformatics, Bologna, Italy, 2002: pp. 99–100.
8. Moreau L, Miles S, Goble et al. On the use of agents in a bioinformatics grid. In: NETTABçAgents in Bioinformatics, Bologna, Italy, 2002.
9. Gonzalez P, Cardenas M, Camacho D, et al. Cellulat: an agent-based intracellular signalling model. BioSystems 2003; 68:171–85.
10. Cássia Trojahn dos Santos, Ana L. C. Bazzan: Integrating Knowledge Through Cooperative Negotiation - A Case Study in Bioinformatics. AIS-ADM 2005: 277-288
11. Bartocci, E., Corradini, F., Merelli, E., & Scortichini, L. (2007). BioWMS : a web-based Workflow Management System for bioinformatics. BMC Bioinformatics, 14, 1-14. doi:10.1186/1471-2105-8-S1-S2
12. Islam, M. T., Bollina, D., Nayak, A., & Ranganathan, S. (2007, March). Intelligent Agent System for Bio-medical Literature Mining. 2007 International Conference on Information and Communication Technology. Ieee. doi:10.1109/ICICT.2007.375342

13. Sophia Ananiadou and John McNaught (editors) ,Text Mining for Biology and Biomedicine University of Manchester and UK National Centre for Text Mining) Boston and London: Artech House, 2006, xi+286 pp; hardbound, ISBN 1-58053-984-X, £53.00.
14. Faro, A., D. Giordano, et al. (2011). "Combining literature text mining with microarray data: advances for system biology modeling." Briefings in bioinformatics.
15. Ana L. C. Bazzan: Agents and Data Mining in Bioinformatics: Joining Data Gathering and Automatic Annotation with Classification and Distributed Clustering. Agents and Data Mining Interaction (2009), pp.3-20
16. Dai, H. J., Y. C. Chang, et al. (2010). "New challenges for biological text-mining in the next decade." Journal of Computer Science and Technology 25(1): 169-179.
17. Ana L. C. Bazzan: Agents and Data Mining in Bioinformatics: Joining Data Gathering and Automatic Annotation with Classification and Distributed Clustering. Agents and Data Mining Interaction (2009), pp.3-20
18. Altman R et al. Text mining for biology — The way forward: Opinions from leading scientists. Genome Biology, 2008, 9(Suppl. 2): S7.
19. Ilić, V. (2009). "Integration of Agents and Data Mining in Interactive Web Environment for Psychometric Diagnostics." Data Mining and Multi-agent Integration: 251-265.
20. Wooldridge, M. J. (2009). An introduction to multiagent systems, Wiley.
21. Faro, A., D. Giordano, et al. (2011). "Combining literature text mining with microarray data: advances for system biology modeling." Briefings in bioinformatics.
22. Popescu, M. and D. Xu (2009). Data mining in biomedicine using ontologies, Artech House Publishers.
23. Smith, L., L. K. Tanabe, et al. (2008). "Overview of BioCreative II gene mention recognition." Genome biology 9(Suppl 2): S2.
24. Cao, L. (2009). Introduction to Agent Mining Interaction and Integration , Data Mining and Multi-agent Integration. L. Cao, Springer US: 3-36.
25. Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A: The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. Nat Biotechnol, 2010, 28(9):897-899.