2011 International Conference on Future Computer Science and Education

Web Search Based on Cultural Algorithm New Framework

Yu Zhang
College of Computer and Information Engineering
Harbin University of Commerce
Harbin,150028,China
zhangyu_20@sohu.com

***Abstract* - With the development of information technology, especially the widespread use of Web, information on Web increases rapidly and becomes a huge information resource. In the meanwhile, such abundant information makes it an urgent problem: how to extract useful content rapidly and efficiently from information resources. This paper proposes a framework for evolutionary systems to search implicit knowledge on the web. Based on Cultural Algorithms (CA), the web search process is supported by the domain knowledge objects in the belief space, and the optimization process is supported by evolutionary search in the population space. This framework in web search may help increase competition and diversity on the web.**

Index Terms* - *Web search, Cultural algorithm, Framework

I. Introduction

The web is unprecedented in many ways: unprecedented in scale, unprecedented in the almost-complete lack of coordination in its creation, and unprecedented in the diversity of backgrounds and motives of its participants. Each of these contributes to making web search different than searching "traditional" documents.

A search engine finds relevant texts containing keywords provided by users based on information search models [1]. However, these keywords are often insufficient and imprecise [2]. This problem results in irrelevant texts being returned and relevant texts being lost. Query expansion improves the descriptive capability of keywords by adding semantically relevant words to original keywords implicitly [3]or explicitly [4]. Expansion words are generated by analyzing their semantic relationships with original keywords. Therefore, web search requires high performance computing, and faces two main challenges: effectiveness and efficiency.

To achieve effectiveness and efficiency, the web search process requires substantial search efforts. This paper tries to apply the power of evolution computation to expedite web search: the search ability of evolutionary computation is used to "select" the necessary cases from a large-scale database to avoid the exhaustive search of every case.

II. The Cultural Algorithm

Traditional Evolutionary Computation methods have limited/implicit mechanisms for representing, storing and transmitting knowledge from one generation to the next. Cultural Algorithm [5], a dual-inheritance evolutionary system, can provide explicit mechanisms for acquisition, storage and refine knowledge obtained during the evolutionary search.

A CA is a knowledge-based evolutionary system. As shown in figure 1, it models two levels of evolution: the population space level and the belief space level. The population space can use any evolutionary population models, such as Genetic Algorithms, Genetic Programming, and Evolutionary Programming [6]. Besides a population space, CA has a belief (knowledge) space in which the problem-solving knowledge can be acquired, reasoned about and refined. The belief space can use any symbolic representation to describe the problem solving knowledge. In this paper, we will develop a more general framework using CA for web search.

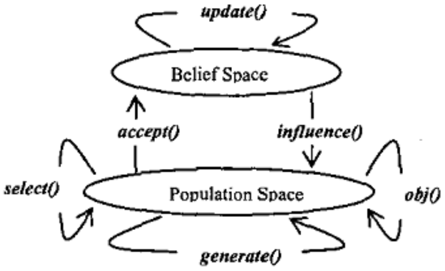

Figure 1 The framework of cultural algorithm

III. The Extended Cultural Algorithm Framework

The Cultural Algorithm framework provides the flexibility of adding knowledge components that are designed for specific problem domains. The framework may he extended to accommodate multiple populations and genetic operators in the population space.

The Canonical Cultural Algorithm is extended to accommodate the new belief and population space objects, as shown in figure 2. Some of the domain knowledge objects in the belief space represent knowledge embodied in a Version Space. This formal structure enables concept learning to occur within the belief space that potentially accelerates member evolution in the population space. Other domain knowledge objects are configured as ontology and taxonomy spaces, and semantic networks.

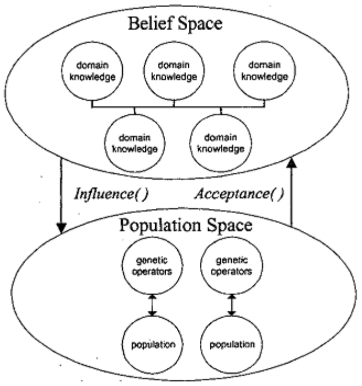

Figure 2 The Extended Cultural Algorithm

978-0-7695-4533-2/11 $26.00 © 2011 IEEE
DOI 10.1109/ICFCSE.2011.141

Ontology may be viewed as an explicit specification of a conceptualization, wherein ontology is a description of the concepts and relationships that can exist between a community of agents [7].

These domain knowledge components within the belief space can be adapted to solve a given class of problems or a particular problem. Heuristics and problem solving operators were recently added to the belief space resulting in a Heuristic Version Space Guided Genetic Algorithm. This addition allows the Cultural Algorithm to contain descriptive knowledge and procedural knowledge [8].

Figure 3 portrays an implementation of the Cultural Algorithm that accommodates web searches.

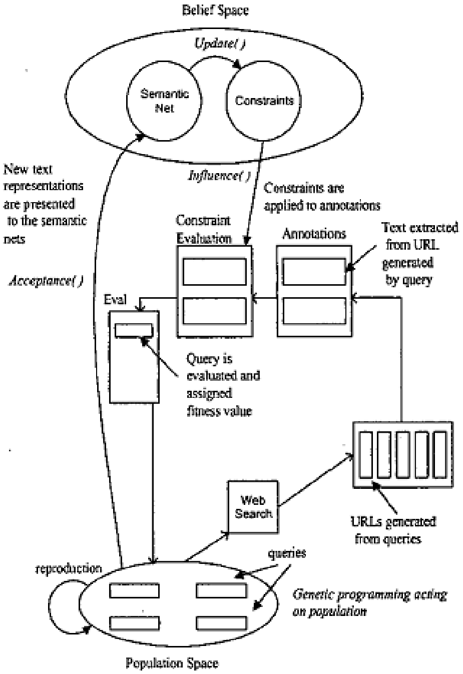
Belief Space
Update( )
Semantic Net
Constraints
New text representations are presented to the semantic nets
Influence( )
Constraints are applied to annotations
Constraint Evaluation
Annotations
Text extracted from URL generated by query
Acceptance( )
Eval
Query is evaluated and assigned fitness value
Web Search
URLs generated from queries
reproduction
queries
Genetic programming acting on population
Population Space

Figure 3 Web Based CA

In this configuration, the population space is based upon a Genetic Programming System (GP), where the population members are tree-structured descriptions of queries [9].

Population members consist of queries to be presented to a search engine. The initial population may be very small, since a single token query will produce many returned URLs and a considerable volume of text when presented to a search engine. After the queries are presented to a search engine, the search engine returns a bag of URLs. Associated with the URLs is text relating to the contents of the web page the URL points to.

Text information is extracted from an indexed Web page and is stored temporarily as an annotation associated with the search query and indexed URL. The annotations are evaluated against criteria provided by constraints residing in the belief space.

The constraints tend to identify and filter noise or 'red herrings' within the annotations. Initially, the population space members may indeed contain a listing of queries producing irrelevant data, or ‘red herrings’. A red herring is text returned by a URL that lexically appears to be meaningful, but whose contents are semantically meaningless.

Domain knowledge components in the belief space contain constraints or rules, which in turn help accelerate the elimination of queries that produce red herrings within the population space.

The tokens and phrases are evaluated by the fitness criteria, assigned fitness values, and are returned to the population space. The population members are operated on by genetic operators or genetic programming techniques, resulting in new search queries. Highly fit queries and selected annotations are also distributed within the belief space objects, which comprise a semantic net, Version Spaces, and a constraint object. Concepts to be learned are placed in Version Spaces.

Members of the population space are also initialized with Web service registry references, gleaned from previous sessions or may be empty. These references may be held in separate populations with separate fitness criteria. In the belief space, the domain knowledge components are assigned their functionality and are initialized accordingly. Version Spaces are initialized by allowing the general layer to accept any concept example and forcing the specific layer to reject any concept example.

Within the repeat loop, the individual populations are evaluated at time t. The evaluation consists of presenting and executing a query to a search engine, extracting text and other URLs from the accessed web page, and submitting the extracted text to the fitness function. Nouns or noun-phrases are extracted from the text as part of the fitness criteria. If constraint elements exist in the belief space, then the constraints are applied to the fitness criteria. This usually resolves to string or token matches, with the constraint object in the belief space containing noun-phrases or tokens.

As the evaluation function terminates, the fitness ranking for the population of queries are determined. Queries that produce red herrings are eliminated from the population or given a low ranking, and relevant queries with respect to the constraint information are assigned higher rankings.

Multiple domain knowledge components exist in the belief space along with multiple populations in the population space. The components within the belief space communicate with each other via a common message-passing scheme.

Normally, Web service business users gather information using search portals and access marketplace applications that interact with web services. Technical users interact with web services directly via UDDI (Universal Description, Discovery and Integration) registries. The Extended Cultural Algorithm can interact directly with and learn from UDDI registries, marketplace and search portals. The diagram in figure 4 shows the Extended Cultural Algorithm’s relationship with marketplaces, search portals, and UDDI registries.

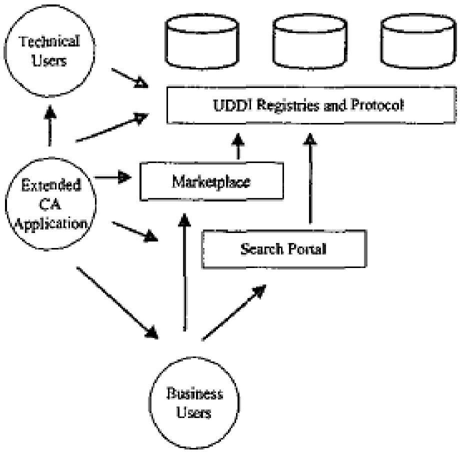

Figure 4 Extended CA Application Space

IV. CONCLUSIONS

This paper proposes a new CA framework to search on the web. In this framework, the data mining that occurs in belief space, and the optimization process that occurs in population space are integrated and reciprocal. This approach suggests a great potential to reach the goal of efficiency and effectiveness for web search.

ACKNOWLEDGEMENTS

This paper is sponsored by the Natural Science Foundation of Heilongjiang Province under Grant No.F200907

REFERENCES

[1] Manning Christopher D, Raghavan Prabhakar, Schutze Hinrich. An introduction to information retrieval. Cambridge: Cambridge University Press, 2008: 109-133; 253-287

[2] Billerbeck Bodo,Zobel Justin.Questioning query expansion: an exam ination ofbehavior and parameters. Proc of the Fifteenth Australasian Database Conference. Dunedin, New Zealand, 2004: 69-76

[3] Cao Guihong, Nie Jianyun, Bai Jing. Integrating word relationships into language models. Proc of the 28th Annual International ACM SIGIR Conference. New York: ACM Press,2005:298-305

[4] Crouch Carolyn J,Yang Bokyung. Experiments in automatic statistical thesaurus construction. Proc of the15th Annual International ACM SIGIR Conference. New York: ACM Press, 1992: 77-88

[5] Reynolds, R. . An introduction to Cultural Algorithms. Proceedings of the 3rd Annual Conference on Evolutionary Programming, Sebald, A.V., Fogel, L. J. Ed. World Scientific Publishing, River Edge, NJ, pp., 131-139

[6] D. B. Fogel, Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. Piscataway, NJ: IEEE Press, 1995.

[7] Gruber, Thomas R., A Translation Approach to Portable Ontology Specifications, Knowledge Systems Laboratory, Technical Report KSL 92-71, Computer Science Department, Stanford University, April 1993.

[8] Stefan, Jeffrey M., Optimal Path Search Using Cultural Algorithms, Masters Thesis, Wayne State University Department of Computer Science, Detroit, MI, 2002

[9] Koza, John R., Genetic programming: on the programming of computers by means of natural selection, MIT Press, 1992