Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No.7, March 2016 – www.caeaccess.org

An Agent-based Meta-Search Engine Architecture for Open Government Datasets Search

S.M. Hasan Mahmud
Daffodil International University (DIU), Bangladesh

Md. Fazle Rabbi
Hajee Mohammad Danesh Science & Technology University, Bangladesh

Kazihise Ntikurako Guy-Fernand
Hohai University, China

ABSTRACT
Recently, Most of the countries are publishing their government data in the Web as datasets. A large number of datasets (HTML, CSV, RDF, XML, JSON, Excel, and PDF etc), catalogs and portals (data.gov, data.gov.uk, data.gov.au etc) are emerging in the Science and Government sector. Open Government Data drives in the US, UK and elsewhere have created hung amounts of government data available to the public on the web. A large number of datasets are published on government data portals, the question arises how to get datasets satisfying from government data portals with the least effort. However in many cases, a single search result is not sufficient to meet the user's query and it is necessary to create a new service platform by combining existing government catalogs search. In this paper we proposed a meta- search architecture that will integrate different dataset catalogs search interface and collected datasets from different open government data catalogs and display higher results. In this proposed architecture have search interface provides a scalable and recognizable solution for finding Open Government datasets from open government data catalogs. Thus, the propose meta-search architecture improves the usability and effectively of open government datasets search.

Keywords
Open Government Data, Government Data Search, Intelligent Agent, Meta-Search Engine, Datasets search.

1. INTRODUCTION
Open Government is a recent act towards more immediacy and legibility in government and governments produce big amounts of datasets as section of their daily activities. Based on the statistics analysis (http://logd.tw.rpi.edu/), the presentence of published dataset on open government web: [Fig. 1] USA (18.5%), Canada (14.9%), Australia (12.3%), UK (9.2%), Spain (8.7%), Italy (7.7%), France (7.2%) and others (21.5%). Open government is be-coming an official policy in several jurisdictions, especially in the US, UK, and Australia. More than 100 government data catalogs are now in operation, listing national, regional while a number of open government dataset catalogs, including Data.gov, Data.gov.uk, United Nations and OpenEI.org have already provided dataset registration and search capabilities. More and more datasets are available on the open government data portal, accessing the right datasets has become a critical problem to search useful datasets. However, users often face problems such as how to find the right and more datasets promptly with the least effort. Search engine is a most prominent and potential tool over the World Wide Web [6].

Generally, search engines search web pages for the queried text or information and then show the outputs by following ranking algorithms [3]. If the user gives a Query as a input, a Search engine displays a record of suitable results arranged in order depend on algorithms [2]. It is the habit of the client to use drop-down approach of the result displayed by the search engine and proves one result at a time, until the required text or data is found. The results on different subtopics or meanings of a query will be mixed together in the list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. A meta-search engine is a search platform that sends user quires to various search engines or databases and accumulates the outputs into a single list or views them corresponding to source. Meta-Search Engine facilitates users to input search query once and access different search engines simultaneously. Here, Our Proposed meta-search engine doesn't have any own database of government data portals. It sends search query to the databases managed by government data Portal search engines and gives users the output that generate from all the search engines queried. An effective and alternate approach to the datasets retrieval on the government data catalogs are by using the meta-search engine, instead of single dataset search platform.

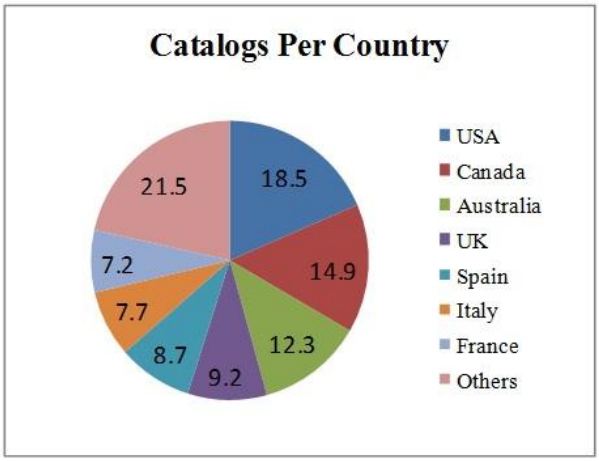

Fig 1: Catalogs per Country

In this paper, we analyze the current Open Government portal internal search engines, justify the need of a new design on how to improve current Open Government search engines, especially a new architecture for Open Government meta-search engines with a multi-agent support in the UI to make search engines work more efficiently and user- amicable [1].

2. PREVOIUS RELATED WORK
Meta-Search Engine is introduced by researchers at an increasing rate. Research efforts in the field of meta-search

Engine have followed several avenues: Early work by Kumar. P [4] has proposed a SEReleC (Meta-search engine) that provided an interface for refining and classifying the search engines results so the displayed search results are more effective. Another similar research work done by Akhlaghian. F and Moradi. P [5] has proposed a multi-agent architecture for personalizing meta-search engine using the fuzzy concept networks. The main work of this meta-search was to use automatic fuzzy concept networks to personalize results of a meta-search engine provided with a multi-agent architecture for searching and quickly retrieving. Bravo-Marquez, F *et al.* [6] have proposed architecture for the retrieval of similar documents from the web. They focused on information search to support the manipulation of users' knowledge into the retrieval algorithm. Their proposed architecture can retrieve similar documents by using the exiting search engines. Raval, V and Kumar, V [10] was present a meta-search engine approach, called EGG (Enhanced Guided Google) that was contracted to use the Google for more authentic and combinatory search. The designed meta-search engine supported the search established on "Combinatorial Keywords" and "Normal Search". Additional work by Chaurasia B, kumar *et al.* [9] deals with priority assist and user profile based meta-search engine. The meta-search engine can improve search act by querying multiple search engines at once. This was to develop for merging the results extracted from two or more search engines. The output and analysis showed that the technique developed the search effectively of the database and fixed search quality was also developed. In [7], the idea of exploiting the scores of each search engine was proposed, where the main information was the relative rank of each result. In [7], different ranking methods were analyzed, for example Borda-fuse which was based on democratic voting, the Borda count, or the weighted borda-fuse, in which search engines were not treated equally [11]. The document retrieving problem has been analyzed by different researchers in [8]. Some approaches were related to fingerprinting techniques for document representation. Also, these approaches used meta-search engine architectures for retrieving an extended list of similar web documents.

3. RELATED TECHNOLOGY

3.1 Intelligent Agent

Intelligent Agents are an emerging technology that is making computer systems faster to use by giving people to surrogate work back to the computer [16]. Today, agents are spared out in different aspect, such as industrial maintain, information searching, personal assistance, network monitoring, games, software distribution, and others more. Agent produces mature standards concerning software agent architectonics and applications and intelligent agents for the searching are specially called information agents [14]. It is suitable of contributing transparent access to one or many different information sources. The key operations such as retrieving, analyzing, manipulating, fusing heterogeneous information as well as visualizing of and controlling the user through the accessible, specific data arena. The agent is able to combine different data and to provide cooperative, multi-dimensional views on similar information to the user [13]. The agent can automatically adjust to changes in user activity, the data, and network setting as well. The main task of agents is to perform searches for, control, and arbitrate similar data on favor of users or agents.

3.2 Meta-Search Engine

The work mechanism of meta-search engine is to sends user query to different search engines and the results into a single list or display according to their source [17]. The member search engines will deal with the actual information retrieval and finally meta-search engine gives users results in a uniform format. So meta-search engine haven't its own information-mining mechanism and database, and its data-collection is based on the retrieval results of member search engines [2]. We use Agent with meta-search engine to come into being effective meta-search engine. There are different ways to improve the performance of search engines. Generally, there are three main directions: I) Developing user interface on query input II) Filtering the query results III) Effective algorithms in web page espionage and collecting, indexing, and output. Our analysis conducted on Internet search engine users and the following features as very important to make a search engine successful: I) User friendly II) Quick Response III) Effective and Accurate Output IV) Update Information date ways. Our study of several Web meta-search engines, we found the following technical decision is most important for meta-search engine: I) How to Parse an input query II) which sources to parse III) How to customize the user imputed query for effective output IV) How to filtering and ordering outputs.

4. AN ARCHITECTUTE OF GOVT DATASET META-SEARCH ENGINE

4.1 Framework for Architecture

This architecture has main three working layers. The first layer is the User Preference Agent (UPA), the middle one is the Information Retrieval Agent (IRA), and the last one is the Result Processing Agent (RPA). The user Preference Agent communicates with users and the architecture: users send the query requests in the layer, the system process the user request by different agent and send the queries to the information retrieval agent [15]. The information retrieval Agent sees to information retrieval: The Agent finds out some search engines fitted into the query requests to retrieve after assaying the demands transmitted [18]. The result processing agent is in responsible of work with the user results: The data filtering Agent filter the retrieve results and the results return to the user preference agent by dataset ordering agents. The architecture figure shows as Fig. 2.

4.2 Architecture Module Achieve

The new architecture proposed is shown in Fig. 2. It is an agent-based approach. It contains a Natural Language Parser (NLP), Query customizer Agent (QCA), Page Retriever Agent (PRA), Datasets Page filter Agent (DPFA), Datasets Ordering Agent (DOA) and User Preference Agent (UPA).

4.2.1 Natural Language Parser (NLP): The Natural Language Parser (NLP) understands the natural language query. Users can input a keywords/query in a natural language to describe the queries or needs. Few search users are not expert to writing the appropriate queries; the parser helps the users to input quires and gives the users to possibility information.

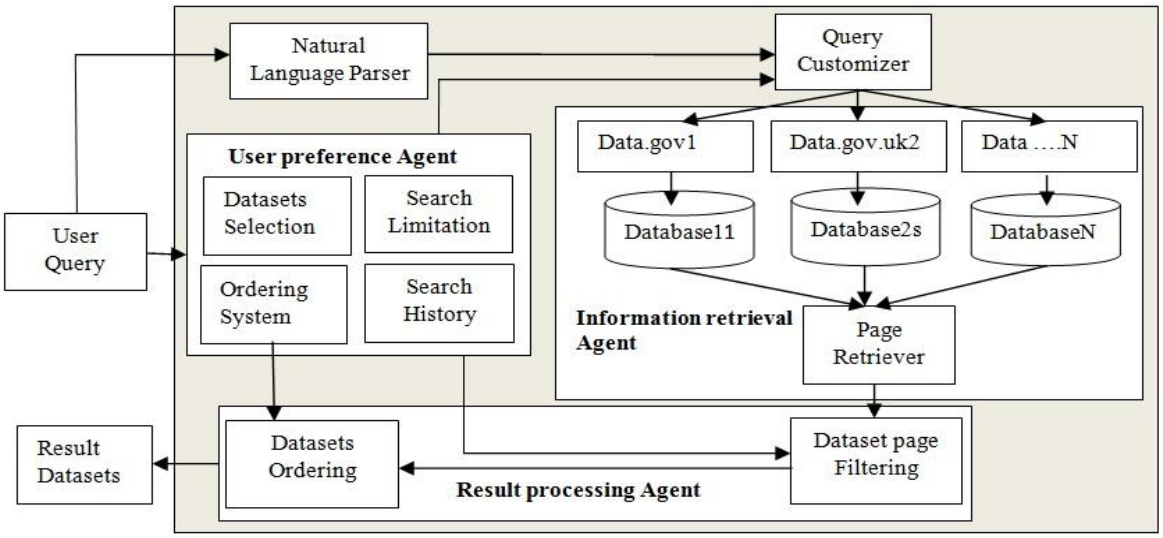

Fig 2: Architecture for Datasets Meta-Search Engine

4.2.2 User preferences Agent (UPA):

User Preferences Agent (UPA) is consisting of Datasets selection, search limitation, ordering system and search history. Users can Change the search preferences at any time to using the User Preferences Agent. Few default modules are provided for non expert users to get the appropriate feedback from simple queries.

4.2.2.1 Datasets selection (DS):

Users can select a search engine to search from a meta-search engine, also user can select the sources datasets format, such as "CSV", "RDF" "XML" etc.; or select the source link from a list of specified sites in the bookmarks.

4.2.2.2 Search Limitation (SL):

Users can manage a search using Search Limitation (SL) module. Our proposed meta-search architecture accept users to modify the query by defining the search range – such as sort by date & time period, language, user areas, also user can define the number of displayed output. In the search range, we allow users to specify which categories they want to search by selecting from a list of pre-defined categories. Users can specify the "type" of the organization they want.

4.2.2.3 Ordering System (OS):

The result datasets ordering system (OS) is the most important part in a meta-search engine, user cans Oder results datasets using Search Dataset Ordering. This architecture accepts to Oder the results by ordering system–The results can be ordered by time, degree of relevancy, or other indexing criteria chosen by users.

4.2.2.4 Search History (SH):

A feedback mechanism is accepted in this architecture. The search results are successfully displayed to the user, the search history is also listed in the user preferences part. A system is running to analyze the arrangement of the using sources of an exact user, and user's performance history, thus show users feedback on their search favorite.

4.2.3 Quarry Customizer Agent (QCA):

Query Customizer Agent (QCA) is used to customize the user queries, analyze the query request according to the relevance input [12]. The selection of the search engine sources, the query modification modules are combined with the Natural Language Parser (NLP) to customized query. The modified query will be sent to the underlying search engines.

4.2.4 Information Retrieval Agent (IRA):

The system process the user request by User preference Agent (UPA) and sends the queries to the search engines using Query customizer Agent (QCA). Query customizer Agent (QCA) send the user request to exiting search engine and the Information Retrieval Agent (IRA) collect result datasets from search engines.

4.2.4.1 Page Retrieval Agent (PRA): An agent called Page Retriever Agent (PRA) will analyze the results from each search engine and pass the results to a page-filtering agent.

4.2.5 Result Processing Agent (RPA):

Result processing Agent process the retrieved datasets that comes from search engines. The Datasets page filtering Agent (DPFA) filter the datasets and send that datasets to Dataset Ordering Agent (DOA) for ordering and display the final result to user.

4.2.5.1 Dataset Page Filtering Agent (DPFA):

The constraints of search range and the sites of search source specifications will form the filtering policy and affect the filtering process. The duplicate and out of range information will be removed by the Dataset Page Filtering Agent (DPFA), the "good" pages will be sent to the Dataset Ordering Agent (DOA).

4.2.5.2 Datasets Ordering Agent (DOA):

How to order the results is one of the most important decisions made by a meta-search engine. By using an Ordering System defined by the User's Preference Agent (UPA), different users, even with the same query and the

same set of documents, will have results presented in an order meaningful to their individual need. The results can be ordered by time, degree of relevancy, or other indexing criteria chosen by users.

4.3 Work-Flow of Meta-Search Engine

Fig. 3 shows us the workflow inside the agent-based search engine. The first search Engine will do is to analyze the keyword input by the user. Our system will analyze these inputs and handle the queries using Natural Language Parse (NPL). Then user can select some different modules (Dataset Selection-DS, Search Limitation-SL, Ordering system-OS, Search History-HS) using User Preference Agent (UPA) and send those quires to the search Engines using Query Customizer (QC) [14]. Page Retriever Agent (PRA) is analyzing the results from each search engine and passes the results to Dataset Page Filtering Agent (DPFA) for filter [13]. The Dataset Page Filtering Agent (DPFA) can remove duplicate results and out of range information and send the result datasets to Dataset Ordering Agent (DOA). Dataset Ordering Agent (DOA) is connected with User preference Agent (UPA), so user can select some module to get more effective result datasets from meta-search Engine. Finally the Dataset ordering Agent (DOA) pass the results set to user for display.

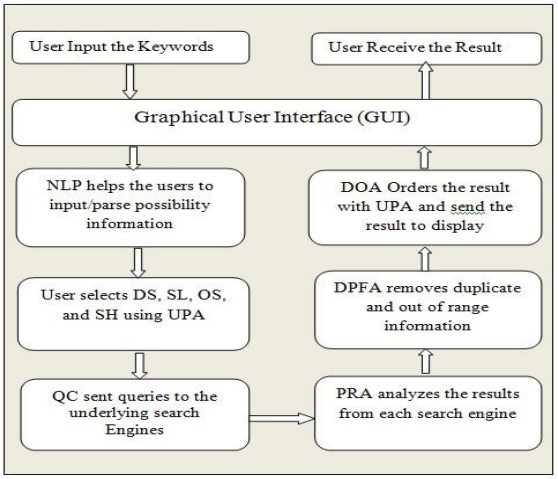

Fig 3: workflow of Meta Search Engine

4.4 Comparing and Analyzing With Existing Meta- Search Engines:

A tabular form is showing the summarizes of seven popular meta-search engines that are developed in past shown in the Table 1. Here, Meta-search Engines are compared by following parameter: Exiting Search Engine**,** Relevance**,** Search engines mentioned, Approximate Amount of Results and Meta Search Topics.

Basically the most of the meta-search engines are using popular search engine (yahoo, bing, google, baidu, ask, altra vasta etc) inside of system but our proposed open government meta-search engine will contain top 50 government search portals (data.gov, data.gov.uk, data.gov.au etc) to get more relevant result datasets. General meta-search engine displays the different web links and documents but this meta-search engine will display the government dataset sets in different format such as CSV, RDF, PDF, XML, Excel, etc only from government portals, so the approximate amount of result will be 80-100% based on user search ability. Most of traditional meta search engine results are unable to fulfill the user requirements and this meta search engine displays results using dataset filtering agent, so this meta-search engine are more effective and accurate then general meat search engine. Our Meta search engine will show high relevance result because of the agent in your system.

Table 1. Summarizes of seven popular meta-search engines

Search Engine/Parameter	Exiting Search Engine	Relevance	Search engines mentioned	Approximate Amount of Results	Meta Search Topics
Dogpile	Google, Yahoo, Ask, Live, AltaVista, DirectHit, LookSmart, FAST, Open Directory ,	High	65%	50-70	Web, usenet, news, ftp
Metacrawler	Google, Yahoo, MSN, Ask, AltaVista, DirectHit, Excite, LookSmart, FAST, Open Directory	Moderate, too many "sponsored results"	82% of engines worked.	50-100	Web, mp3,news, usenet, images
Ixquick	All the Web, Exalead, Qkport, Ask, Gigablast, Wikipedia, Bebo, MSN, Wmzy, CNN, NBC, Yahoo, EntireWeb, OPD	High	39% of engines worked	30-60	Web, mp3, news, images
Mamma	Ask.com, About.com, Entireweb, Business.com, Gigablast, Wisenut, OD	Moderate	60%	60-90	Web, MP3, Images News, Audio, Video.
Dataware	N/A	low	36%	30-50	General, Medical, Financial, Health, Gov.
surfwax	CNN, Yahoo news, HotBot, ODP, Yahoo news, MSN, AllTheWeb	Moderate	40%	50-60	General Web information, PDF, PPT, images etc
Clusty	Ask.com, Gigablast, Live, NY Times, ODP, Shopzilla, Yahoo news, Yahoo stocks	Low, too many "sponsored results	50%	40-50	General Web information, Mp3, Video, images etc

5. CONCLUSION

The Open Government Data catalogs portal are developing in a high speed and datasets are publishing in a large scale, too. The individual Government Data catalogs search engine is not enough to fulfill the user needs, but our proposed meta-search engines architecture will satisfies user's demands. We have proposed an open Government Data meta-semantic-search engine which focuses the idea of integrating search datasets results from existing data government search internal engines such as data.gov, data.gov.uk, data.gov.au etc and display result datasets using result set agents. This architecture provides users with an efficient way to datasets search from different data government search engines by single interface. Our proposed architecture combines search datasets results and brings out only relevant results with help of page retrieve.

6. REFERENCES

[1] Lamberti, F., Sanna, A., Demartini, C.,"A Relation-Based Page Rank Algorithm for Semantic Web Search Engines", IEEE Transactions on Knowledge & Data Engineering, January 2009, vol.21, no. 1, pp. 123-136.

[2] Srinivas, K., Srinivas, P. V. S.,Govardhan, A., "A Survey on the Performance Evaluation of Various Meta Search

Engines", International Journal of Computer Science Issues (IJCSI), May2011, Vol. 8 Issue 3, p359.

[3] Mukhopadhy, D., Sharma, M., Joshi, G., Pagare, T., and Palwe, A., "Experience of Developing a Meta-Semantic Search Engine", International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (2013), p. 167-171.

[4] Kumar, P., " SEReleC (Search Engine Result Refinement and Classification) - a Meta search engine based on combinatorial search and search keyword based link classification", International Conference on Advances in Engineering Science and Management (2012), pp. 627-631.

[5] Akhlaghian, F., Moradi, P., "A Multi-Agent Based Personalized Meta-Search Engine Using Automatic Fuzzy Concept Networks", Third International Conference on Knowledge Discovery and Data Mining (2010), pp 208-211.

[6] Bravo-Marquez, F., L Huillier, G., A. Rios, S., Velasquez Juan D. and Guerrero, Luis A., "DOCODE-Lite: A Meta-Search Engine for Document Similarity Retrieval", In Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems (2010), pp: 93-102.

[7] Rasolofo, Y., Abbaci, F., Savoy, J., "Approaches to collection selection and results merging for distributed information retrieval", In Proceedings of the Tenth International Conference on Information and Knowledge Management (2001), pp.191–198.

[8] Zaka, B., "Empowering plagiarism detection with a web services enabled collaborative network", Journal of Information Science and Engineering (2009), vol. 25, 1391–1403.

[9] Chaurasia, B kumar., Gupta, S Kant., Soni, R., "Meta search engine based on prioritizor", International conference on computational intelligence and communication systems, 2011.

[10] Raval, V., Kumar, P., "EGG (Enhanced Guided Google) - A Meta Search Engine for Combinatorial Keyword Search", Institute of Technology, Nirma University, Ahmedabad –382 481, 08-10 December, 2011.

[11] Aslam, J.A., Montague, M., "Models for metasearch", In Proceedingsof the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001), pp. 276–284.

[12] Chen, J., Liu, W.," A Framework for Intelligent Meta-search Engine Based on Agent", Information Technology and Applications (2005), p 276-279.

[13] Zheng, W., "An intelligent agent-based meta search", International Conference on Future Information Technology and Management Engineering (FITME) (2010), p 263-266.

[14] Klusch, M., "Information agent technology for the Internet: a survey", Data & Knowledge Engineering , Volume 36 Issue 3, March 2001, Pages 337 – 372.

[15] Guoyuan L., Jiutao, T., Chun, W., "Studies and Evaluation on Meta Search Engines", Computer Research and Development (ICCRD), 3rd International Conference on (Volume:3), 2011, P 191-193.

[16] Li, Z., Wang, Y., Oria, V., "A New Architecture for Web Meta-Search Engines", AMCIS 2001Proceedings. Paper 84.s

[17] Sudeepthi1.G., Prof. Surendra Prasad Babu ,M.," Survey On Meta Search Engine in Semantic Web", International Journal of Computer Technology and Applications. 2011;02(06)3051-3055.

[18] MARIAPPAN, A.K., SURESH, R.M., BHARATHI, V.SUBBIAH., "SEMANTIC META SEARCH ENGINE USING SEMANTIC SIMILARITY MEASURE", Journal of Theoretical and Applied Information Technology, 30th April 2013, Vol. 50, No.3.