A COLLABORATIVE, SEMANTIC AND CONTEXT AWARE SEARCH ENGINE

Position Paper

Manuela Angioni, Roberto Demontis, Massimo Deriu, Emanuela De Vita,
Cristian Lai, Ivan Marcialis, Gavino Paddeu, Antonio Pintus,
Andrea Piras, Raffaella Sanna, Alessandro Soro, Franco Tuveri

CRS4 - Center for Advanced Studies, Research and Development in Sardinia, Polaris - Edificio 1, 09010 Pula (CA), Italy
{angioni, demontis, mderiu, emy, clai, ivan, gavino, pintux, piras, raffa, asoro, tuveri}@crs4.it

Keywords: Search Engine, Community, Location Aware, Semantics, NLP, RDHT, DART, 3D-UI.

Abstract: Search engines help people to find information in the largest public knowledge system of the world: the Web. Unfortunately its size makes very complex to discover the right information. The users are faced lots of useless results forcing them to select one by one the most suitable. The new generation of search engines evolve from keyword-based indexing and classification to more sophisticated techniques considering the meaning, the context and the usage of information. We argue about the three key aspects: collaboration, geo-referencing and semantics. Collaboration distributes storage, processing and trust on a world-wide network of nodes running on users' computers, getting rid of bottlenecks and central points of failures. The geo-referencing of catalogued resources allows contextualisation based on user position. Semantic analysis lets to increase the results relevance. In this paper, we expose the studies, the concepts and the solutions of a research project to introduce these three key features in a novel search engine architecture.

1 INTRODUCTION

Nowadays search engines help people to find information but most of proposed results are useless or only partially related to what users need and they are resigned to select by hand the most suitable one. These limits could be overcome refining requests step by step so a real effective could be achieved introducing context and meaning concepts in query composition and resolution. The new generation of search engines requires advanced features and new architectures to mining the deep web (Bergman, 2001) and to find either virtual web objects accessible by browsers and concrete objects or services. Users require information about real objects: available products in a supermarket, a parking close to home, the nearest restaurant, the post office with shorter waiting time. To address such wishes, new search engines should focus on three key aspects: collaboration, geo-referencing and semantics.

Collaboration is the suitable solution to discover the deep web and to distribute the processing power required to scan and catalogue its pieces of information. Peer to peer networks enable users to collaborate submitting directly new resources, offering storage space and bandwidth of their Internet connection. There is no central control system, bottlenecks and central points of failures are avoided and the ranking system is public.

Geo-referenced data management allows users to submit questions related to the position specified by latitude-longitude coordinates or by place names. Search engines will automatically process reverse geo-coding, either during page parsing and user query processing. Mobile objects and people will spontaneously notify their position.

Semantic analysis provides the right meanings of words and sentences according to contiguous text segments and solving misunderstandings related to thesaurus and slang expressions. It improves Web information retrieval, knowledge management and enterprise application integration.

In this paper we expose the studies, the concepts and the solutions developed in the DART project to introduce the three key aspects in a search engine system and relate them with a set of studies.

2 A TOOLKIT FOR SEARCH ENGINES

DART (Distributed ARchitecture Toolkit for search engines) is a research project focused on studying, developing and testing either patterns and integrated tools to improve the quality of search engines results with the main objective to satisfy user needs. Interesting research fields are exploited and merged: semantic-based indexing, P2P crawling, location-aware information retrieval, user requirement filtering, 3D query results interface, virtual assistance and a public Web resources indexing.

Semantic techniques and Natural Language Processing (NLP) tools exceed limits of resources cataloguing and query resolution. An important issue is the designing of more intuitive, adaptable and accessible user interface (UI). In this concern, we are focusing on intelligent virtual assistance and 3D visualization for query responses, to help user during browser session suggesting other resources deduced by current page and current session and filtered using device profile, user preferences and context (i.e. GPS position, weather, and so on). 3D UI lets users to improve their perception of the showed objects thanks to the possibility to explore scenes, move from an object depicting a query result to another, select it and apply rotations, shifts and translations.

Installing a DART Community Node, anyone can join the open and decentralized community. Nodes share computational resources contributing to Web crawling, distributed index storage and acquiring information generated by sensors, GPSs, environmental surveys or events notified directly by users.

3 MODULES OF A DART NODE

The concept of community becomes concrete using the DART Community Node, able to discover and retrieve other nodes and to join them. The node is composed of distinct interconnected modules.

3.1 DDBMS

The Distributed DBMS (DDBMS) is the network overlay specifically designed to meet scalability, fault tolerance, self maintenance and load balancing requirements. It provides a distributed file system called RDHT (Range capable Distributed Hash Table) (Soro et al, 2006), described in §3.2, able to support range queries, and to manage with flexibility, efficiency and robustness a huge variety of distributed applications. Its three primitive operations are: `lookup(value)` to retrieve the entries closest to value or the entry value itself, if it exists; `insert(value)`, to store a new entry in the DDBMS; `range(lb, ub)`, to retrieve all the entries within the interval [lb, ub].

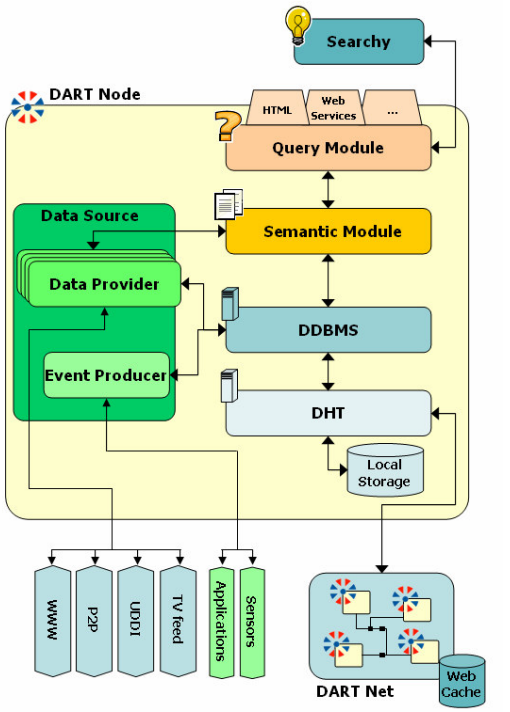

Figure 1: The modules of the DART Community Node.

3.2 Data Sources

Data Sources are represented by two distinct modules with a different retrieval information system: Data Provider and Event Producer.

3.2.1 Data Provider

A Data Provider processes data in order to retrieve all relevant information to send to Semantic Module (see §3.3) for indexing or to be directly stored on the DDBMS. Node contributes to crawl according to hardware configuration and user's preferences. Every node, in base of the optimization policy, can decide to assign crawling of a resource to another node. Nodes that does not crawl can provide other functionalities, such as index storage.

Some P2P bridges retrieve information from the existing resources on the other P2P networks (i.e. Gnutella, eDonkey2K, etc). To collect EPG information a tailored grabber may be installed on iTV platforms like set-top boxes and media centers. Web Services (WSs) are discovered by a module which inquiries UDDI registries.

3.2.2 Event Producer

The Event Producer is a family of software modules able to produce events stored in DART system. An event producer allows to export information about people, objects, and places. The events can be generated by: sensors (RFID, etc.), local applications and peripherals connected to device hosting the DART Community Node. For instance, geographic coordinates collected by a GPS can be added to the resource improving efficacy of the search process.

3.3 Semantic Module

This module is responsible for semantic activities in the system. It works with the DDBMS to create a semantic indexing of ontology-based annotated resources and it supports Data Providers processing texts by means of NLP techniques, interpreting documents, analyzing and resolving queries for the Query Module. Its components are: the Syntactic Disambiguator, a syntactic analyzer that find the syntactical structure of phrases and sentences, in order to resolve roles of terms ambiguity present in natural languages; the Semantic Disambiguator, that identifies synonymy and hypernymy relations from WordNet (WordNet, 2007) and change the representation from words contained in a document to a density function based on the synonyms and hypernyms frequency (Scott, 1998); the Categorizer manages resources and queries classification by means of text categorization techniques; the Semantic Net Manager manages the building of the Semantic Net, composed by a set of topics linked through their senses, and uses the Semantic Net in order to enrich results with topics semantically related to queries submitted by users.

3.4 Query Module

The Query Module collects user queries, processes and forwards them to the Semantic Module. The Query Manager is the part of this module that composes the UI showed to user in base of the information provided by the Device Manager and the User Manager.

The Device Manager provides information useful for an optimal UI rendering according to the device used to submit queries. Moreover it enriches user queries with device features and support results filtering in collaboration with the Result Collector module. The User Manager selects, processes and shares user data with the Query Manager (in query enrichment) and the Result Collector (in results filtering). The Result Collector receives the collection of results from the Semantic Module and processes a more accurate filtering of results thanks to data provided by the User Manager. The Event Manager allows the user to configure a DART Node to produce an alert. The Event Manager is responsible of the user subscription for a particular type of event. Nodes are automatically alerted.

4 NETWORK FILE SYSTEM

DART nodes can store data on a distributed file system called RDHT (Range Capable Distributed Hash Table). The RDHT adapts to the P2P context the algorithms and structures known as Skip Lists (Pugh, 1990), storing in the underlying DHT additional index information, in terms of pointers that link each element to its neighbours. Pointers do not link each element to its next, but simply represent a linear ordering relation that enable to reconstruct the complete list. This solutions is at the same time very easy to maintain and implement, and reasonably efficient.

In addition to low level primitives relative to the DHT operations, that are implemented in the Kademlia protocol (Maymounkov, 2002), the RDHT exposes three primitive operations: `lookup`, `insert` and `range`. `Lookup` operation discovers an item stored in the data structure. Starting from a known value, it executes several `get` operations on the underlying DHT to fetch index information. A recursive algorithm finds the pointers from a known base to the target value. At each step, the algorithm executes the longest possible hop that do not overshoots the target element. This operation fails if the shortest possible hop from every known base overshoots the target, in this case it converges on the elements that are *nearest* to the target.

`Insert` operation executes a `lookup` operation to spot the nearest elements to the target (call them elements A and B); only if the `lookup` fails the new element is stored and pointers from element A to target and from target to element B are stored in the RDHT. Additional pointers are stored to link the

newly inserted entry to other ones in the RDHT. These pointers will speed up `lookup` operations.

Range queries are executed following the shortest pointers from the lower bound to the upper bound of the query. Alternatively a range query can be executed searching for the items nearest to the `median(lb, ub)` and then repeating this operation recursively over the two subintervals until no new element is discovered or a given grain is reached.

Experimental results show that the lookup operation requires the average of O(log N) `get` operations on the underlying DHT, where N is the number of entries in the RDHT. The system is self-balanced in that each node has the same probability to be involved in any lookup operation. The insert operation stores a number of pointers when an entry is added to the RDHT, choosing from a heap of entries, known from previous lookup operations. The goal is to set pointers between entries that have as few pointers as possible, in order to avoid overloading on a single entry. The general criterion to choose a good candidate for a new pointer is:

- entries that are often visited during lookup operations are likely to be overloaded so are bad choices;
- entries that have been recently used have in some ways paid their due so again are bad choices;
- bad choices are pushed to the tail of the heap, thus good choices emerge at the head.

The heap is also used to fetch the base entry for lookup operations. This ensures as far as possible load balancing even when repeatedly polling a lookup over the same value.

5 SEMANTIC ISSUES

Semantics plays an important role in DART pervading either resource indexing or query processing. The semantic indexing over that finding meaning, when it is possible, performs the geo-reference of Web resources. It means to implement the set of algorithms and data models that performs searching and indexing functions of geo-referenced web resources. The Data Providers parsing Web contents identifies structured and unstructured portions. Structured portions are directly indexed by specialized Data Providers using ontology or structure descriptors, as in the case of dynamic XML documents generated by Web services, RSS, GeoRSS and RDF. The last two types of documents have location data defined by name, URL and a Point of Interest (POI) or an Area of Interest (AOI) defined in WGS84 format.

Unstructured contents require linguistic analysis and semantic interpretation, so their processing is delegated by the Data Providers to the Semantic Module, which splits text in phrases to extract "parts of speech" and to identify syntactical relations. Then the semantic analysis finds the right meaning that lets to reduce the false positive in the searching phase. A gazetteer derived from the Geo-Names database (Geo-Names, 2007) is used to locate the possible AOI or POI of the unstructured portion. All POIs are written like a no-area AOI.

The result of semantic indexing is a set of keys that is given to the DDBMS to be indexed. The geographic key is an integer value obtained using a z-curve (Jagadish, 1990) and interleaving the bits of longitude and latitude values. To simplify retrieval operations on geographical areas, the RDHT items related to closest AOIs are directly connected through pointers according to an inclusion area schema defined by a seven-level quad tree representing the Earth. The tree leaves correspond to an AOI of one minute per side.

Regarding to query processing, the Semantic Module analyses the queries produced by the Query Module. Such queries are enriched with context information, user profile and device profile but they need to be again analysed because may be defined in natural language and so it is important to identify the right meaning or in alternative at least to reduce the number of possible meaning to the most suitable ones. An AOI may be directly included in the query by the user or get by the gazetteer. Available AOI is linearized using the same algorithm for indexing.

6 IMPROVING USER EXPERIENCE

In respect to Human Computer Interaction (HCI), we aim at offer user alternative approaches on related GUI aspects supporting user browsing sessions with an Autonomous Interface Agent (AIA) and showing the query result in a 3D environment.

Searchy is an AIA developed like an extension for Mozilla Firefox, that assists and interacts with the user during his browsing session in a not intrusive way. It analyzes the downloaded page and composes a query enriched with additional and relevant keywords according to user and device profile and context information, to submit to Dart Community Node. When the virtual assistant

receives the DART response, it refines and reorganizes the query results thanks to user information, removes banned or useless sites, performs the list ordering and, finally, shows the list in a side bar of Mozilla Firefox. The user profile ranges from personal information to friends, houses, jobs, hobbies, preferences and include data retrieved by browsing sessions such as viewed resources, the time spent on a site, bookmarked pages, etc.

We investigate the concepts of post-WIMP UI improving the user experience thanks a 3D interactive UI presenting data items of query results. It is characterized by concrete representations and simplicity (Houston et al, 2002) with easy to explore scenes and objects that help user to find faster the best information. The 3D-UI module is used according to the user preferences and current user context and transforms a list of results in a 3D interactive scene defined by an X3D document (Web3D, 2007).

7 RELATED WORKS

Methodologies for performing a distributed search and frameworks based on ontologies are developed as a search technology for the Semantic Web (Straccia et al, 2006). A semantic crawler-based indexing and retrieval system is proposed in the SWOOGLE project (Ding et al, 2004), where user searches for concepts instead of keywords.

To improve search engines, we have introduced in DART semantic techniques, such as ontologies for the analysis of structured documents (Saiful Islam et al, 2007), for unstructured documents, we perform a linguistic analysis for syntactical and semantic disambiguation as in (Sleator et al, 1993).

To overcome the typical problems of centralized systems, such as network overload, single point of failure, censorship, lack of scalability, we propose the application of P2P paradigm to realize a distributed search engine architecture. Examples of community oriented architectures for geographic based services are exposed in (Carboni et al, 2006), (Guan et al, 2004) and (MacEachren et al, 2005), while (Callan, 2000) and (Singh et al, 2003) are examples of distributed search engines.

Unstructured systems flood queries to all peers in the network, thus requiring O(N) messages. P-Trees (Crainiceanu et al, 2004) and Prefix Hash Trees (Ramabhadran et al, 2004) are scalable solutions requiring to store a distributed indexing data structure in the P2P network itself, and use this to guide range queries. In order to solve the problem of P2P systems to process range queries, we propose the adoption of RDHT, an overlay adding primitives to manage range queries on the Kademlia.

Internet interfaces of common search engines allow users only basic searches by means of a single search box and a few extra fields in the advanced search mode (Alonso et al, 2007). Presenting results graphically rather than textually, their usability improves and people acquires information more rapidly. Related to HMI interfaces, with 3D UI the user can improve the perception and the understanding of depicted object and, so, of the contents it represents (Biström et al). While our AIA, Searchy, merges several aspects treats in various works, such as an advice system (like (Lieberman, 1995) and (Chen et al, 1998)), the automatic update of user preferences by the analysis of Web-browsing behaviours (Seo et al, 2000) and the use of user profile to complete user requests (Somlo et al, 2003).

8 CONCLUSIONS

Due to the enormous number of digital resources, current search engines show lacks in helping people to find the right information. The DART project aspires to overcome some of this limits integrating solutions related to distributed systems, semantic Web, geo-referencing and HMI. Patterns and technologies have been exploited in the design of a distributed, semantic and context aware search engine. The research effort is directed to provide a toolkit for indexing online resources based on the idea of an open community. Installing a node, people join the community offering storage, processing power and bandwidth. Other advantages are: no central control system exists, bottlenecks avoided, public ranking system and users directly involved in search engine activities.

Numerous studies enrich the community features. We take care on the semantic analyses of structured and unstructured portions of digital resources, user queries may be submitted in natural language and any time it is possible the resources are geo-referenced. The available DHTs have not an indexing system for our analyzed resources so we introduce the RDHT overlay. To support user interaction we work on HMI and personalization. In despite to the most common search engines, we introduce Searchy to assist user during browsing sessions and 3D visual environment to show the query results where he is free to move. A strong personalization is obtained thanks to a combination

of user profile, device features and context of use (i.e. user position). Such combination is used either in query processing to add more details to queries and results selection to exclude false positive items.

Our project is still an on-going process. Although it needs to be tested, rendered more robust and other DHT algorithms and implementations have to be evaluated, it demonstrates it is possible to merge collaboration, geo-referencing and semantics and apply them to a novel search engine system.

The DART research project is partially funded by the Italian Ministry of University and Scientific Research, contract grant number 11582.

REFERENCES

Alonso, O., Drake, M., Banerjee, S., The Information Grid: A Practical Approach to the Semantic Web. Retrieved January 18, 2007, from Oracle Technical Information Web site: http://www.oracle.com/technology/tech/semantic_technologies/pdf/informationgrid_oracle.pdf

Bergman, M. K., 2001, The Deep Web: Surfacing Hidden Value, *Journal of Electronic Publishing*, University of Michigan Press, Vol. 7.

Biström, J., Cogliati, A., Rouhiainen, K., Post- WIMP User Interface Model for 3D Web Applications, Helsinki University of Technology Telecommunications Software and Multimedia Laboratory.

Callan, J., 2000, Distributed information retrieval, *Advances in Information Retrieval*, Kluwer Academic Publishers, pp. 127-150.

Carboni, D., Sanna, S., Zanarini, P., 2006, GeoPix: Image Retrieval on the Geo Web, from Camera Click to Mouse Click. In *MobileHCI'06*, Helsinki - Finland, ACM Press.

Chen, L., Sycara, K., 1998, WebMate: a personal agent for browsing and searching. In *2nd international Conference on Autonomous Agents*, Minneapolis - United States, ACM Press, pp. 132-139.

Crainiceanu, A., et Al, 2004, Querying Peer-to-Peer Networks Using P-Trees. In *WebDB Workshop*.

Ding, L., et al, 2004, Swoogle: a search and metadata engine for the semantic web, in Thirteenth ACM international Conference on information and Knowledge Management, Washington – USA, ACM Press, pp. 652-659.

Geo-Names.org. Retrieved January 18, 2007 from http://www.geonames.org/export/

Guan, J.H., Zhou, S.G., Wang, L.C., Bian, F.L., 2004, Peer to Peer Based GIS Web Services. In *XXth ISPRS Congress*, Istanbul – Turkey.

Houston, B., Jacobson Z., 2002, A Simple 3D Visual Text Retrieval Interface. In *TRO-MP-050 - Multimedia Visualization of Massive Military Datasets*.

Jagadish, H. V., 1990, Linear clustering of objects with multiple attributes. In *SIGMOD'90*, pp. 332–342.

Lieberman, H., 1995, Letizia: An agent that assists web browsing. In *International Joint Conference of Artificial Intelligence*, Montreal - Canada.

Maymounkov, P., Mazières D., 2002, Kademlia: A peer-to-peer information system based on the xor metric. In *IPTPS '01: Revised Papers from the 1st International Workshop on Peer-to-Peer Systems*, Springer-Verlag.

MacEachren, A. M., et al, 2005, Enabling Collaborative Geoinformation Access and Decision-Making Through a Natural, Multimodal Interface. *International Journal of Geographical Information Science*. Vol.19, (19) , pp. 293-317

Pugh, W., 1990, Skip Lists: A probabilistic alternative to Balanced Trees. In *Workshop on Algorithms and Data Structures*.

Ramabhadran, S., et al, 2004, Prefix Hash Tree - An Indexing Data Structure over Distributed Hash Tables.

Saiful Islam, A., et al, M., Ontology for Geographic Information. Retrieved January 18, 2007, from Drexel University Web site: http://loki.cae.drexel.edu/~wbs/ontology/iso-19115.htm

Scott, S., Matwin, S., 1998, Text Classification using WordNet Hypernyms. In *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

Seo, Y., Zhang, B., 2000, Learning user's preferences by analyzing Web-browsing behaviours. In *4th International Conference on Autonomous Agents*, Barcelona - Spain, ACM Press, pp. 381-387.

Singh, A., Srivatsa, M., Liu, L., Miller, T., 2003, Apoidea: A Decentralized Peer-to-Peer Architecture for Crawling the World Wide Web. In *SIGIR 2003 Workshop on Distributed Information Retrieval*, Lecture Notes in Computer Science, Volume 2924.

Sleator, D.D., Temperley, D., 1993, Parsing English with a Link Grammar. In 3*rd International Workshop on Parsing Technologies*.

Somlo, G. L., Howe, A. E., 2003, Using web helper agent profiles in query generation. In *AAMAS '03: 2nd international joint conference on Autonomous agents and multiagent systems*, ACM Press, pp. 812-818.

Soro, A., Lai, C., 2006, Range-capable Distributed Hash Tables. In *3rd International Workshop on Geographic Information Retrieval - GIR'06*, Seattle – USA.

Straccia, U., Troncy, R., 2006, Towards Distributed Information Retrieval in Semantic Web. In *3rd European Semantic Web Conference (ESWC-06)*. Springer Verlag.

WordNet. Retrieved January 18, 2007 from http://wordnet.princeton.edu.

Web 3D Consortium - Overnet. Retrieved January 18, 2007 from http://www.web3d.org.