A Specialized Multi-agent Search Engine Model for the Extended Manufacturing Enterprise

Ockmer L. Oosthuizen, Elizabeth M. Ehlers

Abstract:

The Internet and Internet search engines have revolutionized the way in which we search for, and integrate information into our daily lives. One of the problems with general Internet search engines is the diversity of contexts queries can assume. Therefore, the construction of context-restricted search engines designed for use within a specialized domain has been proposed (S. Lawrence, 2000).

The objective of this article is to suggest a model for a collaborative, personalized meta-search agent system for the virtual manufacturing enterprise in the context of resource location and integration across scope of the extended manufacturing enterprise.

The key differences between the proposed model and general search engines are (1) the ability to personalize the search task according to an individual user need, (2) the utilization of similar search sessions through partner collaboration at an information retrieval level, (3) leveraging on existing search services available on the virtual enterprise extranet and (4) autonomous behaviour through the use of intelligent agents.

The perceived benefits of a such search agent system for the virtual enterprise are: (1) Enabling a consistent and uniform view of business entities through improved access to information, (2) Simplified access to operational data, (3) Effective access and retrieval of organizational model elements and (4) Improving data location and sharing between trading partners providing a platform for tighter process integration between trading partners. These benefits are critical for effective decision making in a collaborative engineering process followed by the differing entities in a virtual manufacturing organization to ultimately enable final product realization.

An evaluation strategy for the COPEMSA system presented is then briefly discussed, metrics for evaluation of focused crawlers and justification for development and implementation of the search system proposed. Finally, this article concludes with remarks about the model presented and future research to be undertaken.

Keywords: *virtual manufacturing enterprises, agent based systems, personalised web search engines*

1. Introduction

The growth of the Internet over recent years has not only enabled easy access to volumes of publicly accessible information globally, but has also substantially contributed to the arsenal of useful, effective and most importantly open technologies available to developers.

One of the most useful applications deployed on the Internet and, more specifically, the World Wide Web (WWW) are so called web search engines. The search engine's primary goal is to effectively locate and retrieve information that is the closest match to a specific user's input query. Advances in Web searching technology have made Web search engines an indispensable tool for the effective use of the WWW.

This article investigates the viability and usefulness of the application of Web search technology to virtual manufacturing enterprises. The ultimate goal of such a proposed search engine is the support of process integration of content between different organizations in the same enterprise and that of trading partners (S. Khoshafian, 2002).

The existence of a collaborative search infrastructure could greatly assist in the integration of existing resources in the extended manufacturing enterprise to improve operational activities and service levels by ensuring that information is easily accessible and readily available to all entities in the organization. The ultimate goal of the organization-wide use of such an infrastructure is the support of effective decision making in the (possibly distributed) engineering process for product or service realization across the virtual manufacturing enterprise.

In the following sections, a brief introduction to the extended manufacturing enterprise and specialized search engines is given. The collaborative, personalized meta-search agent (COPEMSA) system architecture is then introduced. The COPEMSA search system is realized through the use of a multi-agent architecture. Each agent in the search system represents a specialization of the overall search task. We define a user agent, search agent, resource crawler, partner agent and results analysis agent.

The user agent learns about the user in order to introduce personal preference into user queries. The search agent attempts to match these personalized queries it receives from the user agent to multiple resources identified by the Resource Crawler. The Resource Crawler continually navigates the extranet and indexes resource locations and descriptions. The community agent enables the search system to communicate and leverage on the search experiences of other members of the extended manufacturing enterprise. Finally, the results analysis agent organizes and ranks the returned results according to user and partner profiles.

A section on evaluation strategy and evaluation metrics is then presented. The goal of this section is to elaborate on the strategy to be followed to test the proposed search system and also the possible performance metrics that could be used for testing and evaluation purposes. Finally, the article closes with conclusions and further research opportunities.

2. The Extended Manufacturing Enterprise

L. Song and R. Nagi (1997) define a virtual enterprise (VE) as an organization constructed by partners from different companies who collaborate with each other to design and manufacture high quality, customized products. The authors also note that VE's are usually product-orientated; team-collaboration styled and are fast and flexible to respond to changing market conditions. L.M. Camarinha-Matos (1999) define the virtual enterprise paradigm as a temporary alliance of enterprises that collaboratively share skills and resources to competitively respond to business opportunities.

Both the definitions given above, although slightly different, define the following three generic characteristics:

- The virtual enterprise is composed of different entities allied to achieve some goal.
- The virtual enterprise relies on collaboration and communication of information and skills between these entities to achieve its goals.
- The virtual enterprise is highly flexible and adaptive to remain competitive in changing market conditions.

The first characteristic is related to the scope of the virtual enterprise. The question that must be asked here is if these entities are merely different units in the same organization or totally different and independent companies. The second characteristic defines some sort of collaboration between differing entities to achieve goals. The main question of interest here is at what level this collaboration and co-operation is done. Finally, the third characteristic requires that virtual enterprises are adaptable and flexible to changing business needs. The question here is what business model and interactions are needed to achieve this flexibility and adaptability.

These three characteristics and the questions associated with them are briefly discussed in the following subsections.

2.1. The scope and core features of virtual enterprises

S. Khoshafian (2002) summarizes the scope of virtual enterprises into the following three categories: (1) *Process integration between operational units in the same organization*. (2) *Process integration between organizations in the same enterprise*. (3) *Process integration between trading partners*.

One of the characterizing features of the virtual enterprise is the collaboration between entities to achieve goals and to react to changing conditions. It is obvious that this collaboration between entities must be facilitated on various levels throughout the virtual enterprise. S. Khoshafian (2002) defines four aspects or levels of particular importance to virtual enterprises, namely: (1) *A consistent view of business entities*. (2) *Consistent operations on business entities*. (3) *Uniform organizational model*. (4) *Processes with consistent roles and activities VE wide*.

2.2. The virtual enterpise business model

As have been previously mentioned, flexibility and adaptability are two enabling features of the virtual enterprise to respond to changing market conditions. These features are, in part, achieved through the inherent network-structure of the virtual enterprise to enable flexibility and adaptability.

As illustrated in Figure 1, trading partners (TPs) provide different value and support products and/or services to the value chain of the virtual enterprise. This effort ultimately contributes to the realization of a real product/service to the end consumer.

As all resources of trading partners are not necessarily centrally located, it is natural to assume that these trading partners are linked by some form of electronic communications network to enable information exchange and inter-organizational communication (Rautenstrauch, T., 2002). This extended information network between trading partners is commonly referred to as an *extranet*.

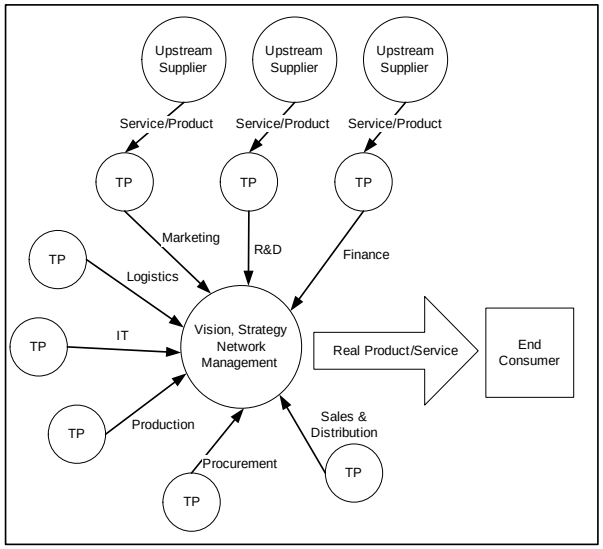

Figure 1. Virtual Enterprise Organizational Model [Adapted from (Rautenstrauch, T., 2002)].

2.3. Information exchange between entities in the virtual enterprise

The final consideration of interest in the context of this section is the type of information exchange between entities in a virtual enterprise.

Virtual enterprises can exchange a variety of information including expertise, product models, business process information, quality-related information, commercial information etc. The only requirement is that some common reference model exists so that all the entities involved in the exchange can participate in the exchange in a pre-defined consistent manner.

Standards like the standard for the exchange of product model data (STEP) and the standard for electronic data interchange for administration, commerce and transport (EDIFACT) are both examples of common reference models for product and commercial information (ISO10303-1:1994, 1994) and (ISO9735-3:2002, 2002) .

The unifying idea between all these standards is to provide a common format for representation and exchange of operational data between collaborating entities.

3. Specialized search engines

The ability to search large volumes of information in a relatively short period of time has been one of the greatest enabling factors of the widespread use of the World

Wide Web in recent years. In this section we briefly discuss the basic ideas behind general web search engines and introduce the concept of a meta-search engine. We then turn our attention to a critical evaluation of current search services and why there is a need for specialization and personalization of the search experience as well as the benefits of context and search space restrictions for search systems. The section closes with a discussion on the benefits of such a context restricted specialized search system for the virtual manufacturing enterprise.

3.1. General web search engines

The Web can be seen as a large collection of highly-topical, geographically dispersed, interlinked information. To enable the effective location and retrieval of information from this very large corpus of information, so called Web search engines were developed. The search process can be summarized in five steps (A. Broder, 2002):

- The search engine user has a certain task he/she wishes to complete and identifies an associated information need.
- The information needed is expressed in terms of a query to be submitted to the search engine.
- The search engine interprets the query.
- It selects those documents that match the query from a corpus of web documents/content.
- Finally, the search engine displays the results of the query to the requesting user (Further refinement of the query based on the results is, of course, also possible at this point).

As a highly successful and effective search engine Google® and (more specifically) Google's desktop search deserves special mention in the context of this paper. The desktop search product is intended as a local search service for the desktop computer with (typically) a single user. This is achieved through building an index of email, files and web history stored on the user's machine and updating it as new information becomes available (Google, 2007). As correctly noted by Chirita *et al.* the desktop search product includes no metadata whatsoever in their system, but just a regular text-based index (Chirita *et al.*, 2005). The collection of metadata on user search sessions and augmenting results based on user context derived from the collected metadata could lead to great improvements in the search experience and perceived usefulness of the search system.

One of the main problems facing many search engines is the issue of equal coverage of the corpus of documents. A related problem from the user perspective is the complexity associated with the use and management of multiple, usually differing search systems. The following section introduces the concept of meta-searching as a technique to address these problems.

3.2. Meta-search engines

Web search engines typically do not offer equal coverage of the World Wide Web (S. Lawrence and C.L. Giles, 1999). In addition to this there are a multitude of search engines available, each offering a different strategy and a different interface for searching the web. To receive the broadest coverage and get more relevant results for a given query, users could use multiple search services.

This can be problematic however, as the overhead involved in collating and merging multiple results can become quite a monolithic task for the user. Moreover, if a large amount of search engines are available, it can also be difficult for a user to keep track of which search services are better for which queries (Glover E. *et al.*, 1999), (S. Lawrence and C.L. Giles, 1998), (E. Selberg and O. Etzioni, 1997).

In order to leverage the coverage of multiple search engines and to provide a single interface to a host of search engines, the concept of a meta-search engine was born. A meta-search engine is a computer program that uses multiple search engines (usually in parallel) to process user queries. The process is briefly summarized in Figure 2 below.

The obvious benefits are that the meta-search engine's coverage is improved through the use of different search engines and that the user is presented with a single interface through which he/she can search the web. Another, less obvious, benefit is that a meta-search engine can be programmed to "know" a lot more about a search engine than a casual human searcher. Many search engines have special features and optimizations that meta-search engines could exploit and take advantage of, thereby improving the results returned from the various search engines (Glover E. *et al.*, 1999), (S. Lawrence and C.L. Giles, 1998), (E. Selberg and O. Etzioni, 1997).

Some search engines like ProFusion (Gauch *et al.*, 1996) try to select the most appropriate search engine(s) for a user query using a learning approach. In essence, the approach keeps track of how well search engines perform for submitted queries of a given context. For each category, n training queries are submitted to the known number of search engines and each engine's performance ranked by some function, indicating how relevant the top x returned documents were to the query submitted. The average scores of the n training queries is calculated to give an overall score for a given engine in a given context. This approach "scores" search engines based on the perceived relevance of documents retrieved for a specific category. These scores are then used to select the most appropriate engine for a query in a given category.

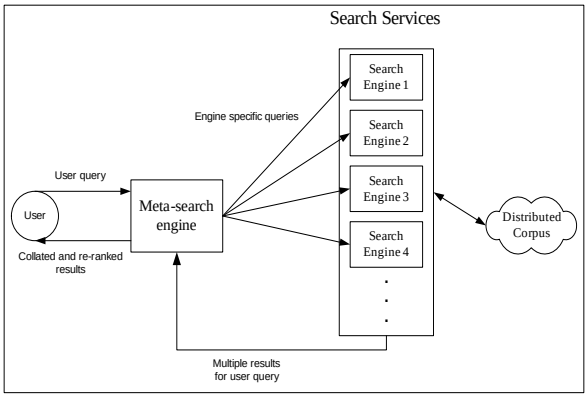

Figure 2. The Meta-Searching Process [Adapted from (Glover E. et al., 1999), (S. Lawrence and C.L. Giles, 1998), (E. Selberg and O. Etzioni, 1997)].

3.3. The need for personalisation and specialisation

Search services are typically used by users with differing goals and informational needs. Currently, the most popular method for the representation of user queries is with a textual string containing keywords. One of the core challenges general search engines are faced with is to determine the context in which keywords contained in a user query was meant. Unfortunately, because of the usually diverse user base and the high volume of requests usually received by such a search engine, personalization is extremely difficult or impossible. The goal of the personalization of search is the construction of a search engine that "knows" their users in terms of the context of their queries, previous requests made to the search engine as well as personal interests and searching habits of the specific user.

Making search engines "context-sensitive" can be achieved in two primary ways:

- Personalization through user modelling.
- Specialization of the search engine to a specific domain or task thereby limiting the context and scope of potential user queries.

These two differing approaches are very briefly discussed below.

3.3.1. PersonaliZation through User Modelling

The idea is to supply the search engine with information pertaining to the user in terms of the user's search habits, frequently used keywords and reaction to presented results. Users can be modelled by either requiring them to complete a user profile before he/she commences to use the search engine or the system can dynamically build a user profile through observation of the user's interaction with the search engine.

3.3.2. Search Engine SpecialiZation

Context sensitivity can also be achieved by restricting the search engine's scope, effectively focusing specialization of the engine to a particular domain or context. This has the benefit of simplifying the interpretation and processing of keyword-based user queries as the number of contexts a specific keyword could have been meant in is automatically limited by the restricted domain.

Another benefit of specialization is a potentially smaller corpus that the search engine has to cover. This could enable the engine to do a more comprehensive analysis on the corpus documents without a significant time trade off.

Furthermore, as noted by Poblete and Baeza-Yates (2006), the application of text-mining and link analysis techniques to the analysis of the content distribution and link structure of a website could lead to improvements in relevant content provision as well as interconnectivity between similar contents. The application of models for improving the distribution of content and link structure like the one suggested would benefit search engines by simplifying content location and restricting the number of links to be followed for retrieval.

3.4. Specialized search for the virtual manufacturing enterprise

In the context of the virtual manufacturing enterprise, specialized and personalized search engine technology could improve communication, collaboration and interoperation between entities in the enterprise.

As was noted in the previous sections, one of the enabling requirements for the virtual enterprise is the existence of a digital computer network between entities. This implies that users located at a certain entity have access to information and services located on the networks of other entities (i.e. departments, organizations or trading partners).

It is also not uncommon for companies deploy applications and content on their intranets using Web technologies. As an example, consider a manufacturing enterprise that publishes and maintains a comprehensive parts list and ordering system on their Intranet. This parts inventory and ordering system should then also be available to any trading partners the enterprise might have through the network link that exists between the two organizations.

The value proposition posed by search engine technology is the ability to locate and retrieve information about products/services/processes published on the organizational extranet by partner entities. Additionally, the improved access to information helps stimulate trust and improve communication between entities in the virtual enterprise, both which are critical components for effective decision making in a collaborative engineering process followed by different entities for final product realization.

The ability to easily search and retrieve information on partner extranets helps to enable the four core virtual enterprise features (see Section 2.2).

- Search capability enables easy access to information held by partners, thereby promoting a consistent view of business entities.
- Searching exposed operational databases on a partner's intranet can help ensure consistent operations on business entities and provide a convenient method for extracting audit trails.
- The search system can help promote a uniform organizational model by enabling access to documents pertaining to the corporate structure of a trading partner.
- Finally, a search system deployed on the virtual enterprise extranet can help ensure that processes are understood to have consistent roles and activities by partner entities by granting access to policy and procedural documentation stored on the trading partner's intranet.

To support the abovementioned, the search system can be specialized according to the user's organizational role, thereby automatically limiting the context of submitted queries. It would also be advantageous if the search system were to operate *autonomously*, freeing up the user's time to pursue other activities.

In the following section, we present a search system model to achieve context-sensitive query formulation and refinement as well as autonomous operation.

4. THE Collaborative, PersonaliZed META-SEARCH AGENT (COPEMSA) SYSTEM ARCHITECTURE

The remainder of this article focuses on the development of a collaborative personalized meta-search agent

system for the virtual manufacturing enterprise. In section 4.1 the design goals of the system are outlined as well as the benefits and limitations of the chosen approach. Section 4.2 presents the COPEMSA system model and outlines key technical details of the model.

4.1. COPEMSA System Design Goals

The potentially vast amount of information available on the organizational extranet could make it one of the largest sources of information available to the virtual organization. Unfortunately, because of the part structured, part semi-structured nature of the information on the extranet; it can be a difficult and extremely time consuming task to locate and retrieve relevant and useful information. As more information and services become available on extranets between partners, it is of paramount importance that methods are developed to aid in the easy recovery of this information.

As was mentioned in section 3.3, users typically differ in their information needs and have differing interests and characteristics. Personalization of partner search could lead to more accurate query formulation and thereby increase the relevance of returned results. The idea of a community of people, perhaps located at different organizations in the virtual enterprise, with similar interests could also aid in improving search results and exchange of search experiences.

One of the key issues to be addressed by any personalized information recovery system for the virtual organization is the issue of coverage and is discussed in the following subsection.

4.1.1. Issue of coverage

Users would obviously want the maximum coverage of the extended enterprise possible when their search queries are processed. The only currently feasible method for achieving this is to attempt to leverage on search services already available on partner organizations extranets, thus leveraging search coverage. Meta-searching is therefore also a key concept in the design of the system.

Intelligent agents act autonomously on behalf of their users. This is desirable for the search system because of potential time savings. A multi-agent system could autonomously continue searching and evaluating extranet content, thereby automating the information retrieval task to a certain extent. Web content mining techniques could also be successfully applied to analyze retrieved pages from the partner's extranet for potentially useful information. Web mining techniques could also be applied to the results retrieved from meta-searching the extranet. Users could then focus more of their time on evaluation of retrieved results instead of exhaustively searching for information.

Based on the discussion above, the critical evaluation of current search services and the need for specialization and personalization given in section 3, the scope of development for the collaborative personalized meta-search agent system can be summarized in the following 5 points:

- Personalization of search through user modelling.
- Collaborative filtering of search engine results and partner-sensitive searching.
- Meta-searching the extranet for improved coverage.
- Using multiple intelligent agents for autonomous and continuous extranet searching and result evaluation.
- Use of Web content mining techniques for analysis of multiple search engine results and content published on the organizational extranet and/or the World Wide Web.

A further design issue is the notion of **where** the search mechanism should be located. There are two possible approaches to this, either on the **client side** or **server side** and is discussed in section 4.1.2.

4.1.2. Issue of Location of Search Mechanism

Conventional search engines typically follow the server side model, i.e. user queries are processed remotely by a server. Large scale personalization of a server based search service could potentially be extremely expensive in terms of processing. Keeping track of a large user base's previous queries, selected documents and other personalization information could be an extremely difficult and processing intensive task, slowing down search engine query processing speed as a side-effect.

A client-based approach could therefore be more feasible for a personalized search agent system. This is discussed in more detail below.

There are a number of significant benefits in a client-side approach to personalized searching. Such a system could more effectively monitor the behaviour of a user, thereby building a more reliable user profile. A client-based system could then effectively modify user queries to help retrieve documents that are relevant to a given context. Another benefit, in the context of a collaborative multi-agent system, is that the processing of retrieved content could be distributed across multiple system users in the same organization or partners, thereby reducing the load that would normally be on a server or group of servers in a server based approach. This information could then be shared among members of the virtual enterprise in an attempt to minimize reprocessing of results.

One of the limitations of a client-side approach is that such a service would not have local access to a large scale index of the content available on the entire extranet, thereby limiting their functionality (S. Lawrence, 2000). As mentioned before, the use of meta-search techniques could improve this limitation significantly.

4.2. The collaborative personalized meta-search agent (copemsa) system model

In this section, a model is given for the COPEMSA system. The system's key components are then briefly discussed.

The model given in Figure 3 is a client-based multi-agent system for meta-searching the extended manufacturing enterprise. The system also provides for collaboration between multiple users.

The system consists of five key components: role agent, query agent, partner agent, result analysis agent and a directed extranet crawler.

The above core components are discussed in more detail below.

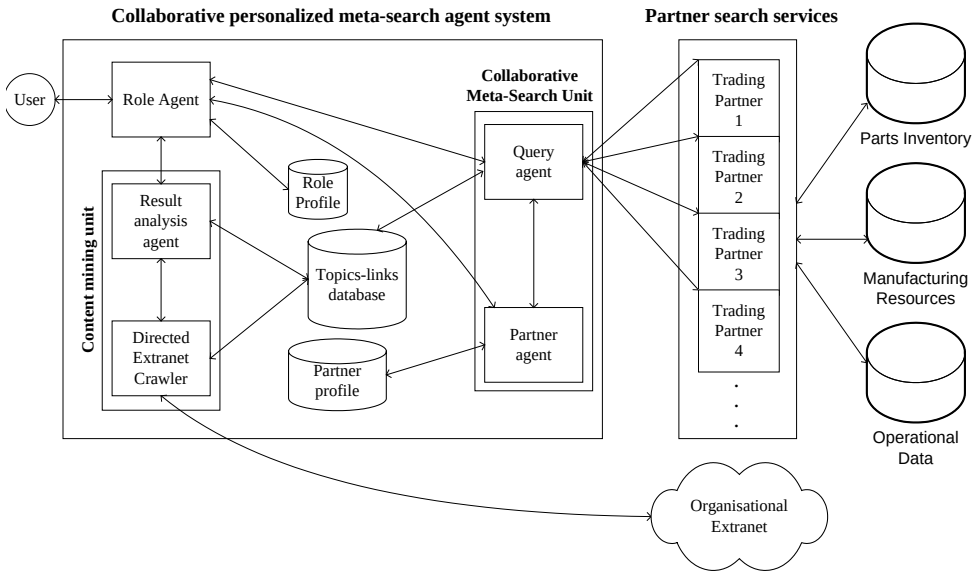

Figure 3. The COPEMSA System Model [Adapted from (O. L. Oosthuizen, 2004)].

4.2.1. Role agent

The role agent is responsible for information collection and modelling of its user(s). It stores information related to the individual user's role in the organization in a *role profile database*. It is important to note that this database can contain the role profiles for multiple roles, either for the same user or multiple users as a search system like this could potentially be used by multiple users of a given organization.

The role agent receives search query strings from the user and then uses the user profile database to rewrite the query to include context specific terms. The modified query is then passed on to a *query agent* for further processing.

The profile agent achieves this by maintaining a *keyword hierarchy* for each profile as illustrated in Figure 4.

The structure of the keyword hierarchy tree is based on that of the Open Directory Project (http://rdf.dmoz.org/). The role agent monitors the queries submitted to the search system as well as the user/role's reaction to the returned results. The agent then modifies the keyword hierarchy based on its observations. The internal nodes represent topics with further contextual specialization as the tree is traversed downwards. The leaf nodes represent specific keywords associated with a specific context. Additional information stored at leaf nodes about keywords are their usage frequency and their perceived relevance to the specific role being profiled.

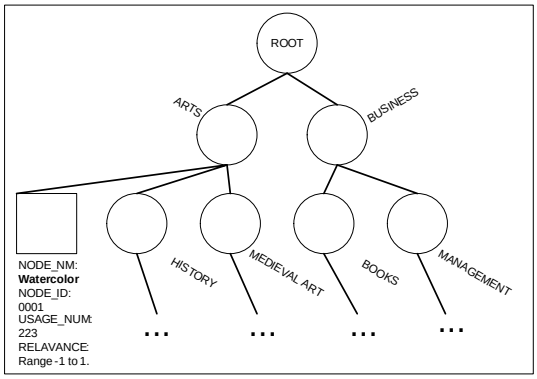

Figure 4. ODP Node Annotation Approach [Adapted from (O.L. Oosthuizen, 2004)].

The process described above can be seen as consisting of three crucial steps, namely:

- **Building a user profile for each user represented as an annotated ODP-tree in the system and using this profile in query augmentations.**
- **Storing the role-profile in a database for each user using an individualized ODP tree storing descriptive words/phrases about contexts defined in the general ODP tree**. The descriptive words stored in the individual ODP tree are words that define a given context for a specific user. This implies that, although all individual trees have a similar structure, the words stored at the leaf nodes may differ for each user of the system.
- **Using user feedback to refine the agent's perception of its user**. Implicit and explicit feedback can both be used to refine the role agent's perceptions about its user's interests. This means that the role agent interacts with the results analysis agent to determine what keywords/key phrases were present in a given result. The ODP-tree node annotation scheme relies on keywords/key phrases to represent user interests. By combining user ratings (either explicitly or implicitly) with an analysis of the most frequently used keywords/key phrases in a document may be an indicator of how good (or bad) a certain keyword/key phrase is for describing a context in the ODP tree structure. Through the feedback process, the user agent can modify the relevance indicators stored at each leaf node in his/her individual ODP tree. This constitutes the refining of the agent's belief about its user in the model presented in this section.

A concern that must be addressed in the functioning of the role agent as described above is the possibility of the agent modifying queries and thereby filtering documents of interest to the user due to an underdeveloped role-profile stored in the role profile database for that specific user. This situation is avoided by pre-training the user agent with initial interests and keywords frequently utilised by the user. This ensures that the agent is able to

rewrite queries to obtain documents that are more likely to be perceived useful to the user from initial use.

4.2.2. Collaborative meta-search unit

In the context of the virtual enterprise the ideas of a community or alliance between organizations (and implicitly also between individual roles or members of the organization) and shared resources across the enterprise between different entities play a critical part in the operation of the organization.

Collaboration among these roles is necessary to ensure the efficient operation of the virtual enterprise as a whole. The promotion of access to shared resources not only helps to streamline the enterprise's processes, but also promotes improved relationships between entities by enabling transparency of data and operations.

The collaborative meta-search unit in the COPEMSA system consists of two agents to incorporate collaboration between search users and integration of shared resources. The *query agent* processes user queries and the *partner agent* is responsible for interfacing with other search system users. These will be briefly summarized below.

Query agent

The query agent is a specialized meta-search engine. It is responsible for the processing of modified queries. This involves rewriting the queries in a search-service specific form (if applicable) and then posting them to a host of different organizational search services (if any are available). Organizational search services could include services like searchable parts inventory, manufacturing resources published on a partner's intranet or a repository of operational data such as transaction logs etc to name a few.

One of the main challenges facing the query agent is the rewriting of the user/role query in a search service specific way. This is accomplished in the COMPEMSA system through 1) the use of the role profile maintained by the role agent and 2) the use of an *interface definition* for search services.

Queries are represented internally in the system through an XML mark-up scheme. The actual query string is typically coupled with information such as the query type, node ID etc. The query is then augmented by using the role profile maintained by the role agent. This includes the contexts the keywords in the query have been associated with and any additional keywords associated with the identified contexts.

Next, the query agent uses an interface definition to guide the transformation of the annotated query for specific inter-organizational search services. These interface definitions contain information about what kind of queries the search service can process. For example, does the service support search phrases, forced exclusion of keywords, forced inclusion of keywords, etc. In the context of a virtual organization, the interface definitions for search services deployed on the extranet can quite easily be agreed upon and shared between trading partners.

Using the annotated query and interface definition, the query agent generates a number of different queries and submits them to the identified search services, effectively meta-searching the organizational extranet for relevant results.

After the various search services have responded to the submitted queries, the query agent is also responsible for receiving and collating multiple search service results (i.e. removing duplicates, ranking of results etc.).

The results received from this process are then stored in a *topics-links database* that lists the topics a user or role is interested in and the locations identified by the query agent to potentially be within that topic.

PARTNER agent

The partner agent initiates contact with other multi-agent systems of other users located either in the same organization or another entity outside of the organization. The agent acts as an information broker between the role and query agents and the virtual enterprise's search community. The partner agent supplies additional community-sensitive information to the other agents on demand and facilitates the exchange of user sessions and analyzed results between members of the virtual enterprise's search community.

In the COPEMSA system, this is done through the use of a centralized partner server. The primary focus of the partner server is to provide multiple instances (potentially extended enterprise wide) of the search system with additional keywords and contexts. The server is responsible for storing and providing access to information submitted by the various partners to aid each other in their searching activities.

The partner agent periodically collects all the local ODP-trees located with other users of the search system for personalizing user queries from the user agent and submits these to the partner server. This action forms the basis for building a *partner ODP* tree that consists of common words for describing a particular context. The structure of the partner tree is similar to that of the individual ODP trees maintained by the user agents of the various members of the search community. In order to factor out the words most commonly used to describe a given topic/context by the community, the partner agents submit each individual user's local ODP-tree to the server for processing. On the partner server, the ODP-trees submitted by the agents of different community members are analysed and the most common words associated by community members with certain topics are included in the partner ODP-tree as common words.

User agents may also submit a topic/context to the server for lookup in the ODP-Tree. The topic is matched in the partner ODP-tree with an internal node of the same name, and any common words (if any) is returned to the requesting user agent. These can then be used by the search partner's user agents for query augmentation.

The use of a central server has the advantage of simplifying agent discovery and communication. Instead of taxing every system's partner agent with the task of automatic discovery of other users using the same search system, the agents could simply communicate through a central server thereby ensuring higher participation and information submission by partners. The additional contexts received from partners is then passed to the query agent to assist in query formulation or used to modify resource rankings directly in the topics-links database.

4.2.3. Content mining unit

The type of content published on the extranet that exists between entities in the virtual organization could possibly be of a highly diverse in terms of context and structured, unstructured or semi-structured in terms of organization. The content mining unit consists of two components, the *results analysis agent* and a *directed extranet crawler* to enable the search system to effectively deal with these two challenges.

The results analysis agent continually monitors the topics-links database for new content to analyze. The agent performs a post-retrieval analysis on the content to enable improved ranking and presentation of results.

The directed extranet crawler is responsible for retrieval of actual content (e.g. data from operational databases, content published on the extranet as web pages etc.) from the organizational extranet, as directed by the results analysis agent.

These two agents will be briefly elaborated in the sections below.

Results analysis agent

The results analysis agent uses web content mining techniques to initially analyze results retrieved from the query agent to determine the context and topic(s) of the page. The process the results analysis agent follows to achieve this consists of three phases: identification, analysis and ranking.

In the identification phase the agent identifies potential pages from the topics-links database and then uses a *directed extranet crawler* to retrieve information from that page or resource directly. The benefit of direct retrieval and analysis is the customization of the ranking of results for various roles based on the role profile maintained by the system. Post-retrieval analysis of results can also assist in the presentation of the results in a role-specific way.

In the analysis phase, the agent mines the content of the identified resources to further its knowledge about the resource. The approach taken by the agent to analyze the identified resources is a process consisting of the following steps:

- **Document representation and feature extraction**. A standard "bag-of-words" (D. Mladenic,1999) approach could be utilized where an individual document, d_i is then represented as a vector of features d where each feature is associated with a word frequency (i.e. the number of times the specific word appears in the document). One of the major issues with the use of word-based features from resources is the problem of the resultant document vectors being of very high dimensionality. This may have a serious impact on the computational effort required for processing large numbers of vectors in a relatively small space of time. A common approach is to remove words that occur in a stop list from the feature vector. The stop list typically contains words that are common to the language most retrieved results will be written in. Other, more advanced techniques like latent semantic indexing (LSI) with singular value decomposition (SVD) have also been applied to the dimensionality reduction problem (S. Chakrabarti,2003).
- **Neighboring resources**. Careful analysis of anchor-text in hyperlinked resources could prove invaluable to additional resource discovery by the results analysis agent. An issue that must be addressed is the problem of resources with a deep and dense link structure. These resources could have many circular references to each other and could force the agent to analyze the same pages repeatedly. The agent could address this by only considering results that are a preset distance from the original page, thereby ensuring that identical resources are not exhaustively retrieved and analyzed.
- **Document classification and Clustering**. In this step the agent attempts to automatically discover general patterns present across the multiple results and group the results according to similarities. A common approach is term frequency inverse document frequency (TFIDF) document classification. The TFIDF scheme represents a key classification method and variants of it are used by many content mining and meta-search systems (D. Mladenic, 1999), (D. Dreilinger and A.E. Howe, 1997). Document clustering techniques can be used to group similar documents into clusters, with each cluster representing a topic or subtopic. This idea of automatically grouping similar documents is of key importance to the results analysis agent as it would then be able to automatically create a taxonomy of results grouped by topic. The clustered results could then be presented to the user in a much more structured way and/or the generated taxonomy could be used as the basis for ranking and filtering operations on the result set or even further query specialization for submission to the query agent (E. Han, *et al*, 1998).

In the context of the virtual manufacturing enterprise, the structure of the information shared between trading partners is usually well defined (see section 2.3). The context of the information received is also typically restricted to the activities of the manufacturing enterprise. Leveraging on these two assumptions in the analysis phase could greatly increase the quality of the returned results in terms of relevance to the user query.

In the final ranking phase, the agent modifies the topics-links database with the newly discovered knowledge gained (from the clustering and classification operations discussed above) about the resource(s). After the clustering and classification process, a taxonomy of resources classified into different clusters or topics will have been generated. The agent then associates the newly discovered knowledge with the various topics defined in the database. This information is then used in the following ways by the search system:

- Future query augmentations by the query agent.
- Results presentation through clustering by topic.
- Further Role-profile specialization.

The results analysis agent could naturally be customized to suit the needs of different manufacturing enterprises and the different informational needs of the user base.

Directed EXTRANET CRAWLER

The directed extranet crawler is a small scale web spider capable of retrieving text-oriented content and link-

following. It has the key goal of retrieving pages and data from the organizational extranet for analysis by the results analysis agent.

A general web crawler (also known as a web spider) can be defined as a software program that automatically traverses the web by downloading documents using the standard Hypertext Transfer Protocol (HTTP) and following hypertext links inside these documents to other documents. General web crawlers usually traverse the web on a large scale, i.e. they download a wide variety of web pages, process the information contained inside these pages and follow hypertext links inside the processed page to a linked page where the process repeats itself.

A special class of web crawlers called *focused crawlers* exists where the goal is not to download pages and follow links on the massive scale general web spiders do, but rather to only gather documents and information on a specific topic or from a specific resource. Focused crawlers typically use less network bandwidth and download fewer items when compared to general spiders as only items on a certain topic or from a certain resource is retrieved.

The directed extranet crawler in the COPEMSA system is a focused spider that collects information from specific sources on the organizational extranet related to some topic. The spider is controlled by the results analysis agent and requests pages as described for analysis.

As noted by (S. Mukherjea, 2000), the main bottleneck for a web crawlers is the time spent downloading documents. In order to speed up the results analysis process, the directed spider could fetch all the documents associated with a specific query in parallel and store the HTML source of the documents or extracted database information in a local cache for fast retrieval by the results analysis agent. Additionally, by only considering topically related information the download of irrelevant pages is avoided.

5. Copemsa system evaluation strategy and metrics

With the discussion of the COPEMSA system given above, attention can now be given to the questions of perceived efficiency of COPEMSA compared to other web crawling software, and evaluation metrics that can be used for the evaluation and performance testing of web crawling software and finally if the development and implementation of COPEMSA is justified in terms of the given metrics.

5.1. COPEMSA EFFICIENCY

One of the development goals of the COPEMSA system was the ability to meta-search the organizational extranet for improved coverage (see Section 4.1). COPEMSA also utilizes a focused crawler for retrieval of topical pages related to user queries, if needed. This guarantees that COPEMSA is at least as efficient as the underlying search services it uses. Furthermore, because post-retrieval analysis is done on results by the results-analysis agent based on a personal profile, COPEMSA is potentially better at satisfying the informational need driving a specific user's queries than more generalized crawlers.

5.2. Evaluation metrics

Two classic metrics for the evaluation of the efficiency of a retrieval system exist in the information retrieval literature, namely precision and recall. *Precision* refers to the proportion of relevant documents retrieved from all documents in the corpus. *Recall* refers to the proportion of relevant documents retrieved from all relevant documents available in the corpus.

In their paper, P. Srinivasan *et al.* (2002) define a general framework for evaluation topical crawlers. They utilize four measures: precision and recall for target pages that have been identified by the specific topic and precision and recall for relevance assessments based on the lexical similarity between crawled pages and predefined topic descriptions (P. Srinivasan *et al.*, 2002).

The directed web crawler component of the COPEMSA system presented in this article can be evaluated using a similar approach and generalized metrics as presented by P. Srinivasan *et al.* The approach can also be extended to evaluate the precision and recall for the search system as a whole.

A third metric, *user perceived relevance*, can also be integrated into the performance metrics for the COPEMSA system to gauge the users perception of the results returned by the search system. Through interaction with the user agent, users rate and score keywords/key phrases as described in section 4.2.1. The annotated ODP-tree maintained by the user agent can then be used as the basis for determining the perceived relevance of search terms through the usage number and/or relevance modifiers associated with each term.

Further analysis can additionally be done using the usage number and relevance modifiers using clustering techniques to generate clusters of keywords/key phrases against which the system can be gauged.

5.3. Development justification and advancement relative to the current state of the art

Based on the COPEMSA system development goals stated in section 4.1, the main justification for development of the COPEMSA system in terms of the two metrics stated in the previous section are improved precision through the use of meta-searching techniques by the query agent and improved recall through post-retrieval analysis and re-ranking of results through the use of web-content mining techniques by the results-analysis agent.

Models like the one proposed in this paper represent only one of many steps toward the ultimate goal of enabling users to construct individual views of the web according to their own personal criteria. More specifically the model adds value in the following areas:

- *Content location and personalization of searching*. Even with the use of general search engines, finding useful information on a large, interconnected network is still a tedious time-consuming effort. Search systems that autonomously locate documents on behalf of users and then proceed to analyze and categorize the located documents based on personal preferences represent a significant enhancement to current generalized search solutions.
- *Collaborative rating of content*. The inclusion of a partner agent in the search system model could

prove invaluable for determining the (perceived) quality of content deployed on the intranet and the design of ranking algorithms that take these ratings into account when ranking results. Content with a perceived low quality can then be gradually phased out of the result sets presented to users.

- *The issue of coverage*. The query agent presented in the proposed model leverage on the combined efforts of various search services provided on a corporate extranet. The meta-search ability of the query agent contributes significant value to the search system due to improved coverage of contents as well as the ability to interface with each individual search service in an optimized manner.

6. The next generation internet (NGI) and Internet2

Internet2 is made up of a consortium of leading U.S. universities working in partnership with industry and the U.S. government's next generation internet (NGI) initiative for the development of a faster, more reliable Internet for research and education. Included in this initiative is the development of enhanced, high-performance networking services and the advanced applications that are enabled by those services (Katz *et al.*, 2001). One of the more interesting features of Internet2 is its planned inherent support for middleware. This layer of software provides core services and the idea of core services becoming part of the networking infrastructure is of prime importance in the context of this paper. It can be argued that these core services provided by the networking infrastructure should include resource location and retrieval services. The integration of models similar to COPEMSA into the set of core services offered by such a networking infrastructure could lead to improved resource location and personalised results presentation to the final user.

7. Conclusions and future research

In this article we discussed the scope, core features and business model behind the virtual manufacturing enterprise. We also discussed the ideas behind specialized search engines, meta-searching and personalization of the search experience.

The main focus of the article was the introduction of the COPEMSA search system architecture for the virtual manufacturing enterprise. COPEMSA is a multi-agent system designed for the autonomous location and retrieval of information on an organizational extranet in a context sensitive manner.

The key benefits of a common search system integrating resources between different organizations in the same enterprise and the resources of trading partners is higher access to information on an enterprise level and increased transparency of operations between trading partners, stimulating trust.

Further research will be focused on the improvement and refinement of the search model specifically for the extended manufacturing enterprise domain. Specifically, practical implementation considerations for the model proposed in this paper for a concrete manufacturing enterprise will receive special attention. The key aspect to success of this system in a manufacturing enterprise is the adoption of the system not only by departments internal to the enterprise but also trading partners. Through leveraging on the inherent trust that exists between trading partners, resources external to the enterprise could be integrated into the results retrieved by the system, thereby delivering results that are more accurate and useful for various users' enterprise wide.

AUTHORS

Ockmer L. Oosthuizen* – Academy for Information Technology, University of Johannesburg, Auckland Park Campus, South-Africa.
E-mail: oloosthuizen@uj.ac.za.
Elizabeth M. Ehlers – Academy for Information Technology, University of Johannesburg
Auckland Park Campus, South-Africa.
E-mail: emehlers@uj.ac.za.
*Corresponding author

References

[1] A. Broder, "A Taxonomy of Web Search". SIGIR Forum, vol. 36, 2002, no.2, pages 3 – 10.
[2] Camarinha-Matos L.M., "Introduction to Virtual Enterprises", Uninova, MonteCaparica, 1999. Available online at: http://www.uninova.pt/~escn/ttt_portugal.html
[3] Chakrabarti S., *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Publishers, 2003.
[4] Chirita P.-A., Costache S., Nejdl W., and Paiu R., "Semantically Enhanced Searching and Ranking on the Desktop". In: *Proceedings of the International Semantic Web Conference Workshop on The Semantic Desktop – Next Generation Personal Information Management and Collaboration Infrastructure*, ISWC, Galway, Ireland, November 2005.
[5] Dreilinger D. and Howe A.E. Experiences with Selecting Search Engines Using Metasearch. ACM Transactions on Information Systems, vol. 15, 1997, no.3, pages 195–222.
[6] Gauch S., Wang G., and Gomez M,. "ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines", *Journal of Universal Computer Science*, vol. 2, no. 9, 1996, pages 637–649.
[7] Glover E., Lawrence S., Gordon M.D., Birmingham W., and Giles C.L., "Recommending Web Documents Based on User Preferences". In: *SIGIR 99 Workshop on Recommender Systems*, Berkeley, CA. 1999.
[8] Google, *Google Desktop Features*, Available at http://desktop.google.com/features.html, Accessed 18/10/2007.
[9] Han E., Boley D., Gini M., Gross R., Hastings K., Karypis G., Kumar V., Mobasher B., and Moore J., "Webace: a web agent for document categorization and exploration". In: *Proceedings of the Second International Conference on Autonomous Agents*, ACM Press, 1998, pages 408–415.

[10] Khoshafian S., "Web Services and Virtual Enterprises", Tect, 2002, Available online at: http://www.webservicesarchitect.com/content/articles/khoshafian01.asp
[11] ISO 10303-1:1994, *Industrial automation systems and integration Product data representation and exchange - Overview and Fundamental Principles*, International Standard, ISO TC184/SC4, 1994.
[12] ISO 9735-3:2002, *Electronic data interchange for administration, commerce and transport (EDIFACT) -- Application level syntax rules*, International Standard, ISO TC154, 2002.
[13] Kratz M., Ackerman M., Hanss T. and Corbato S., "NGI and Internet2: Accelerating the Creation of Tomorrow's Internet". In: V. Patel *et al.*, editor, *MEDINFO 2001*, Amsterdam, 2001. IMIA, IOS Press.
[14] Lawrence S. and Giles C.L., "Inquirus, the NECI meta Search Engine". In: *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia. Elsevier Science. 1998, pp. 95–105.
[15] Lawrence S. and Giles C.L., "Accessibility of information on the web", *Nature*, vol. 400, 1999, pp. 107-109.
[16] Lawrence S., "Context in Web Search", *IEEE Data Engineering Bulletin*, vol. 23, issue 3, 2000, pp. 25-32.
[17] Mladenic D., "Text-learning and related Intelligent Agents: a Survey", *IEEE Intelligent Systems*, vol.14, 1999, no. 4, pp. 44–54.
[18] Mukherjea S., "WTMS: A system for collecting and analysing topic-specific web information", In: *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands, 15th-19th May, 2000.
[19] Oosthuizen O.L., "A Multi-Agent Collaborative Personalised Web Mining System Model", MSc Dissertation, University of Johannesburg, 2004.
[20] Poblete B., Baeza-Yates R., "A content and structure Website mining model". In: *Proceedings of the 15th International Conference on the World Wide Web*, Edinburgh, Scotland, 23-26 May, 2006 957-958, 2006. Review date: 4 Oct 2006. Review published with ACM Computing Reviews.
[21] Rautenstrauch T., *The Virtual Corporation: A Strategic option for Small and Medium Enterprises (SME'S)*, Association for Small Business & Entrepreneurship Annual Conference, 2002.
[22] E. Selberg and O. Etzioni., "The MetaCrawler architecture for resource aggregation on the Web", *IEEE Expert*, January-February, 1997, pages 11–14.
[23] Srinivasan P., Pant G., and Menczer F., "A general evaluation framework for topical crawlers", *IEEE Trans. on Knowledge and Data Engineering*, Submitted, 2002.
[24] Song, L. and Nagi R., "Design and Implementation of a Virtual Information System for Agile Manufacturing," *IIE Transactions on Design and Manufacturing, Special issue on Agile Manufacturing*, vol. 29, 1997, no. 10, pp. 839-857.