



INDIANA UNIVERSITY

SCHOOL OF INFORMATICS

Bloomington

6S: A Collaborative Web Search Network

Prepared for: IBM UIMA Innovation Award

Prepared by: Filippo Menczer, PhD

August 31, 2007

Proposal number:



Executive Summary

Abstract

6S is a collaborative peer network application, aimed to extend the current model of centralized search engines with large numbers of autonomous, distributed, micro-search engines. Each peer within the 6S network crawls the Web in a focused way, guided by its user's information context. This way better contextual coverage can be achieved. Each peer also acts within the network by submitting, forwarding, and responding to queries to/from its neighbors. Peers depend on a local adaptive routing algorithm to dynamically change the topology of the peer network and search for the best neighbors to answer their queries. If realized, the 6S network will lead to an "intelligent network" where peers learn continually about users and other peers, thus addressing user search needs in a contextual, personalized, and scalable way.

Experimental plan

The initial phase of this project was funded by a NSF Career grant. This has proven the viability of the 6S idea via large-scale simulation, and has led to the development of a first working prototype. At the current stage, each peer is identical in functionality and differs only in the focus of its collection; all peers employ the same keyword interface to retrieve and rank results, and the same learning algorithm to track other peers and route queries. Now we propose to deploy such application in the real world and allow for the development of peers with different focus, expertise, and semantic capabilities. Our goal is for each peer to be able to easily implement different algorithms for topical Web crawling, ranking of search results, modeling of other peers, and routing queries based on their semantics. This will enable a new level of community search that may lead to seamless integration of diverse and novel search applications.

Potential benefits from and to UIMA

The proposed network can leverage a much broader community than is possible with today's search engines, even those employing social collaboration. Anyone can develop a search service for any concept, from topical/vertical search engines, to specialized semantic Web services such as travel agents, to universal search—major search engines like Google and Yahoo could participate, and probably would quickly become hubs.

We will explore the use of UIMA to enable this vision, from facilitating the design and development of search services (crawling, semantic retrieval, result ranking) to managing the social search network (learning about other peers, maintaining neighbor profiles, routing queries). We will study how UIMA could be used easily by a community of people building specialized search services, and how the social search network would discover and incorporate the semantic capabilities of new search services that the community comes up with. This may require the development of tools to make UIMA accessible to a broad community of people, not just highly skilled programmers.

Finally we will study how to integrate UIMA with text analysis (indexing, retrieval, ranking) algorithms provided by different open source projects, such as Apache Lucene/Nutch, that are employed in the current 6S prototype.



Project Description

Background and objectives

The Web is used today as a means to integrate many electronic information resources. In particular, it enables the creation of personalized and collaborative services. Since the inception of the Web, however, we have also witnessed a change in the way users seek knowledge and retrieve information. This has been a change from the browsing (a.k.a. *surfing*) behavior inherent in the original distributed architecture of hypertext, towards using the Web as a large database to be queried by powerful and effective search engines.

It is rather difficult for centralized search engines to cover the entire Web,¹ because it is large, fast-growing and fast-changing.^{2,3,4,5} Further, various biases introduced to address the needs of the “average” user imply diminished effectiveness in satisfying many atypical search needs. Examples of bias include interfaces (advanced search features are often buried and poorly documented), ranking (in favor of precision and popularity^{6,7}), and coverage (well connected pages are easy for a crawler to find and thus more likely to be indexed⁸). In spite of enormous progress in crawling performance,^{9,10} indexing,¹¹ retrieval and ranking,^{12,13} the “one engine fits all” model does not—cannot—scale well with the size, dynamics, and heterogeneity of the Web and its users.

The human-machine interaction of existing centralized search methods is particularly rigid: such services do not proactively push relevant information to users about related topics that they may be unaware of, there is typically no mechanism for collaboration among users, and there is also no direct mechanism to exchange data between users. The centralized search model treats users as independent information seekers disconnected from social context, and information resources as passive repositories of data.

It is thus apparent that a central weakness in the current centralized model of search is the lack of *context awareness*. Most search engines and Web portals currently offer customized interfaces, however, they do not provide users with the context-aware personalization and collaborative environments which are more natural for distributed architectures. Namely, the ability to automatically adapt to the information needs of users, and to discover and maintain appropriate communities of users.

In addition to the ubiquitous use of the Web, two recent technological developments make the study of distributed approaches (to supplement existing centralized search) very appealing, feasible, and indeed quite needed: *desktop search* tools and *peer-to-peer* (P2P) networks. It is easy to foresee a near future when users make portions of the data stored in their computers available to others via the Internet. Communities and users evolve, requiring the search environment to adapt. It is therefore imperative to develop distributed algorithms to cater to context-aware, adaptive, community semantics. Naturally, many issues of privacy, security, and copyright arise in such a scenario. Its full potential can only be realized if users are able to, securely, select which subsets of their local information they choose to make available to whom, and under which conditions.

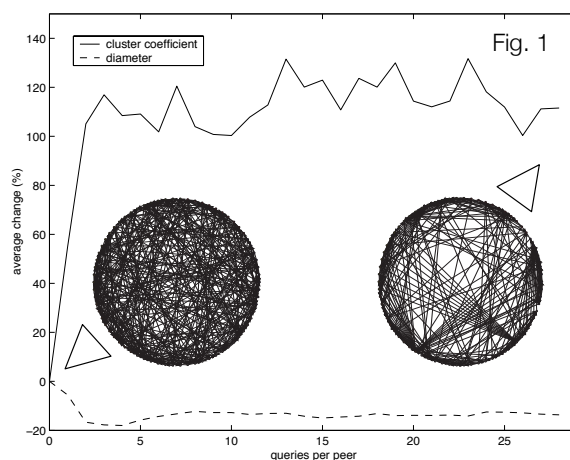


Existing search methods are superb in finding relevant information independently from context. As some search functionalities move to the desktop we see an opportunity to preserve and leverage the diversity of user communities, which is not realized by centralized search. At the same time, research on P2P networks has advanced the production of robust architectures which are ideal to broker and cater to individual and community needs. In this proposal we aim to dovetail these elements to produce collaborative information services capable of better servicing the diverse search and recommendation needs of users. This investigation must and will go hand in hand with an understanding of possible abuses, whether relating to privacy, security, or pure functionality. As an example of the latter, reputation systems are notoriously difficult to build and maintain, and almost always open up vulnerabilities due to their distributed functionality. More specifically, the global functionality depends on locally generated and maintained reputation scores, thus allowing for an adversarial modification of functionality by the synthesis of accounts and reputations. However security and privacy aspects are not the focus of the present proposal and will not be further discussed here.

Our proposed research envisions a federated system of adaptive agents, interacting over a peer network. While agents adapt their strategies based on local context (be it the user's present needs or the local network of peers), the network as a whole adapts its topology. We note here that ours is a distributed approach perfectly *complementary* to centralized systems, which can always serve as "super-peers" in the network.

The 6S framework

In preliminary work the PI's group has begun to explore the main ideas proposed here: a distributed, collaborative approach to Web search designed to provide scalable and robust coverage, while balancing the bias of any single participating search engine. In this view, context awareness and personalization result from an adaptive search interface that learns user preferences. In our envisioned peer network/application, called 6Search (6S), each peer employs a topical crawler to harvest Web pages in a focused way, guided by its user's information context. Further, each peer submits and responds to queries to/from its neighbors. This search process has no centralized bottleneck. Peers depend on local adaptive routing algorithms to dynamically change the topology of the peer network and search for the best neighbors to answer their queries. In preliminary experiments we have evaluated very simple machine learning techniques for local adaptive routing. Simulations with 70-500 model users based on actual Web crawls yielded encouraging preliminary results. By analyzing the dynamic overlay network formed by the passing of query and result messages among peers, we find that the topology rapidly converges from a random to a small-world network (as illustrated by the low diameter and high clustering coefficient in Fig. 1), with clusters emerging to match user communities with shared interests.^{14,15} Additionally the quality of the results is better than obtained by centralized search engines built with equivalent resources, and comparable with larger search engines.^{15,16,17}





These preliminary results suggest that a distributed, collaborative search framework can draw advantages from the context and coverage of the peer collective. The present proposal aims to bring these ideas to fruition, and to extend them by designing a framework in which each peer search engine can easily implement different ranking and learning algorithms.

We envision a federated system of adaptive agents, communicating and interacting over a peer network. Each agent corresponds to a user or set of users, a host machine, and an application working as client and server software. We use the terms *peer* or *agent* to refer to any or all of these components; the meaning will be clear from context. A peer user base may range from a single user to a large community. The host may be a single personal computer, or a large cluster. The application may provide a small service (e.g. searching through a small set of documents on a local disk), a large one (e.g. a commercial search engine), a specialized service (e.g. searching for car rental information), or no service at all (e.g. a user who just queries the network). The proposed distributed model is not meant to replace, but rather to extend the current state of the art, by incorporating centralized, universal search engines as well as smaller topical ones. We expect that commercial search engines may play a key role in such a network, with a mutually beneficial relationship: search engines will provide a critical service as “reference hubs” for many general search needs, and will be able to use 6S as a vehicle for their product according to their business models—for example serving targeted advertising.

Fig. 2 illustrates a peer within the proposed 6S network. The salient features of 6S are:

- The system is *distributed*: information and knowledge are scattered across multiple peers. There can be redundancy; for example important documents and pages should be indexed by many peers.
- The system is *collaborative*: peers exchange information (queries, results) over the network.
- The peers are *context-aware*: The peer's search and recommendation behavior depends on a given user persona, as well as on the local peer community relevant for a specific query and/or persona.
- The peers are *adaptive*: a peer stores models of other peers by learning from its interactions with them. For example, a peer's history of responses over time reveals its expertise and reliability. Likewise, a peer's history of queries reveals its information needs and interests. Such historical data is available not only to the end peers receiving queries and responses, but also to intermediate peers along the paths that propagate these messages. A peer's models evolve over time (e.g. by learning new keywords or documents), increasing accuracy and tracking changes in interests and expertise of other peers.
- The network itself is *adaptive*: Based on the learned models of individual peers, messages are routed (queries and responses, whether generated locally or forwarded) in a smart way. In the case of intermediaries, this is equivalent to a form of *referral*.¹⁸ This leads to a dynamic overlay network topology that is able to capture semantic locality among groups of peers.

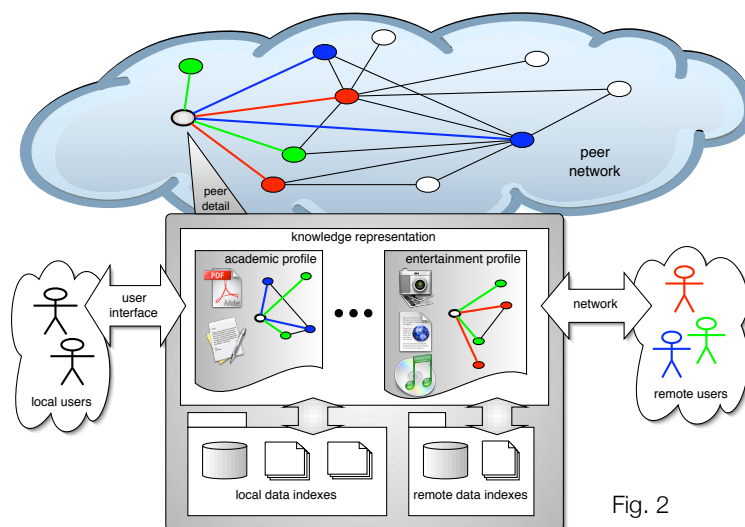


Fig. 2

Research on 6S will lead directly to improved coverage, scalability, and context awareness with



respect to the current model of centralized and universal search engines. Many pages and sites are currently not indexed by search engines, because they are not linked or poorly linked and thus hard for a crawler to reach. However, these sites may be known to some (small set of) users who find them relevant. As such sites are indexed by some peers, they become discoverable by network searches initiated by neighbor peers. As the Web grows and changes, an adaptive peer network can grow and change with it, while preserving the *diversity* of local communities with interests that may be too specific and with too small a link popularity to be picked up by universal search engines. Indeed, in 6S, the diverse resources of the collective can scale up to the challenging tasks of crawling, indexing and retrieval. The fact that a search engine gives low priorities to certain sites because they are not “important” on a global scale is no longer a problem; those sites will be picked up by other peers as long as someone cares.

Finally, each peer has access to a large amount of contextual information in order to personalize and adapt its service. In 6S, contextual specificity is established from two main sources: the present user and its relevant *semantic* neighborhood. The latter context is defined by the other peers that are well known and with whom frequent interactions occur regarding specific sets of topics. In Fig. 2, the interests of a user are represented by icons (e.g., documents related to academic work or files related to music) and its relevant semantic network is represented as a snapshot of (a portion of) 6S. Colors of other peers and connections represent learned associations between the interests of local users and those of neighbor peers.

Peers can employ data mining techniques for topical crawling, indexing, retrieval and ranking that exploit appropriate weighting factors for their various users. As a trivial example, pages in the *edu* domain might be given higher prestige by peers interested in science. Such contextual information could be captured by a personalized version of PageRank.¹⁹ Other examples include the discounting of internal links as nepotistic and the weighting of keywords. The vector representation of page content might be tuned based on the term frequency distributions appropriate for a specific area, discounting vocabulary terms that are too rare or too common in the context of the peer. HTML markup, anchor text, and other emphasis features may also be contextualized by each peer.

There are additional advantages to the proposed approach. One is *diversity*: if many peers contribute to the results of a search, the hit set has a higher chance of representing diverse points of view than if a single, homogeneous method is used for crawling, retrieval and ranking. Moreover each peer has local control of at least part of the process, with the possibility of adjusting the search dynamically. For example a user might exclude, include, prioritize, or randomize other peers to affect the diversity of the results. Another additional advantage of 6S is *robustness*: any number of peers can go down and the network will continue to function — in the limit the network should revert to the current centralized search model. This resilience or graceful failure must be supported by appropriate communication and data mining algorithms, carefully analyzed, and evaluated in the field based on the emergent overlay topologies of the peer network. Other types of robustness, e.g. vis-a-vis malicious peers, will also be studied in depth but are outside the present proposal's scope.



Proposed UIMA related research

Turning the above vision into reality makes it necessary to solve several important problems. Among the many research challenges in the project, let us discuss a few that are particularly amenable to leveraging and contributing to the UIMA platform.

Peer data gathering

Efficient and effective data gathering algorithms are needed; these include intelligent topical crawlers as well as appropriate interfaces for the user to identify the local information (i.e. folders, files and URLs) that can be shared with other peers.

With respect to crawling external data sources, we will build upon extensive prior work by the PI and his group on topical Web crawlers.^{20,21,22,23} One key synergy that the proposed project will allow us to explore, derives from the rich local context available on a peer via desktop search, which can be harnessed for guiding a crawler and helping it focus on the most relevant topics. The profiles from users dynamically track user interests and identify semantically related peers. Rich representations (such as manually curated ontologies, document vectors, associative networks, and even higher-order latent concepts) can be learned and/or evolved by a peer to identify key features and guide crawlers. For example sufficient information will be available to train a crawler in a supervised fashion, or to build sophisticated on-line reinforcement learning models. Such crawlers will effectively harvest highly relevant pages from the Web and make them available to the peer community. Furthermore, we will expand our work into the mutual relationship between crawling algorithms and peer search (indexing, retrieval) subsystems.²⁴ We expect a synergistic effect from having information flow not only from the former to the latter, as in current search engines, but also back from the search subsystem to the crawler, which can use the knowledge inferred from query interactions with local and remote users as a guide. All of these research directions offer opportunities to leverage UIMA. Each peer can use UIMA functionalities to support context-sensitive crawlers and to enable users to transparently guide intelligent crawlers.

With respect to locally indexed data, we have thus far relied on the Apache Lucene and Nutch open source libraries paired with simple in-house interfaces to allow users to manage their local collections. We will integrate the text analysis (indexing, retrieval, ranking) algorithms in these libraries with those in UIMA, and augment the 6S interface with more powerful, scalable, and usable document management and analytics functionalities provided by UIMA. Although well established text indexing techniques and recent developments in desktop search applications provide tools for parsing documents of many types, and for indexing them based on various text and metadata features, there are a few research directions that UIMA will enable. First, we need effective user interfaces to make it easy—even transparent—for users to select the data sources (folders, files, URLs, single documents) that they are willing to share with the peer communities associated with their profiles. Second, a weakness of current desktop search applications is the inability to exploit one of the most important features that allows centralized search engines to rank results effectively: links. We need to associate link information with non-hypertext documents so that peers can locally estimate prestige based on link analysis. To this end we will develop algorithms to automatically link local documents based on content similarity as well as associations extracted from profiles. We will explore the use of UIMA's text analytics, annotation, and social networking features for attacking this problem effectively. However, if



anyone is to be able to use UIMA for deploying a 6S peer for any desired search service, UIMA must be made accessible to non-specialized programmers.

Query semantics

UIMA can be used to support query systems beyond the keyword based interface of the current 6S prototype. For example, suppose that a user submit as a query a question requiring a specific answer. Current research in Question Answering systems highlights the need of NLP techniques to analyze question queries and match them with text that might answer them. For example, questions that start with “when” may be matched with temporal information such as dates. If a peer is capable of answer questions in some domain, then 6S needs to be able to route queries in that domain to the appropriate peer. We will experiment with the application of UIMA to build semantic query analysis components and incorporate them into the 6S query routing algorithms.

Integration of query results

Algorithms must be developed for the integration of peer results. We cannot expect, nor would it be desirable, that each peer conforms to a universal scoring/ranking function. On the contrary, allowing each peer to develop its own ranking algorithm based on local context is just one of the strength of the 6S approach that we hope UIMA will facilitate. How to combine hits then? Some work in this direction has been done in the context of meta-search. One approach is to combine ranks rather than scores. When additional information about the results is scarce, the ranking may be unreliable. Fortunately, peers will learn about other peers from experience, building models of trust, reputation, and reliability. This information will be valuable for a peer to weigh and more accurately rank the combined results. These algorithms must be adapted to the situation in which peers may employ a very large number of different scoring functions with widely different statistical distributions, and to the possibility that some spammer peers will maliciously inject misleading, inconsistent and even adversarial information into the network.

Peer representation and acquisition algorithms

We will use UIMA to design flexible peer representations allowing agents to adapt their strategies (e.g., query routing or keyword selection) according to users and appropriate semantic neighborhoods. The development of robust text/data mining algorithms enabling peers to learn and track each other’s representations from local traffic is key to the viability of 6S. Data will be bursty, noisy, streaming, unstructured or semi-structured, possibly annotated with (inconsistent) scoring.

We have framed the 6S as a technological confluence of P2P networks, desktop search, and adaptive algorithms. A key aspect of this model is the realization that users can be more than generic, passive information seekers, and indeed function as information providers themselves. Additionally, in our vision, the information that users provide, as well as the information they seek, serves to define their distinct contexts. Therefore, we must treat each user as an *information resource* in a distributed architecture of peers. To achieve this, we design peers to represent both the information a user stores and the information a user has sought.

The search strategy of each peer (e.g. query routing and keywords or phrases employed) *adapts* to its context (i.e. peer and community). As peers interact in a network, they learn new information about their appropriate peer



community and about topics of interest. This includes learning about peer trust, reputation and reliability, as well as learning novel keywords and information resources for relevant topics of interest.

UIMA will allow us to explore powerful adaptive representations that peers can use for themselves, for other peers with whom they interact, and for their local network communities. Given such a peer representation, several research problems arise. We will investigate the best methods to automatically learn the “best” acquaintances on the basis of past experiences of interaction. The problem is to determine which peers, among the many with which an agent interacts, should be chosen as the members of a finite acquaintance list, so as to maximize the search performance of the agent. The choice may need to be made dynamically, based on context and on the current information need; different peers may be best at answering different queries, even if they originate from the same user.

In preliminary work we have experimented with simple term vectors to represent peers and simple reinforcement learning algorithms for adaptive query routing. The exploration of alternative and potentially better representations and learning algorithms will be greatly facilitated by the UIM architecture, and in turn we can contribute new representations and algorithms to the UIMA community, to facilitate the development of diverse distributed collaborative search techniques.

Community identification

Finally, given our distributed vision of a peer network, several research problems arise from the need to discover semantically related neighbors and identify communities using only locally available information. As peers exchange queries and facilitate searches, they learn how to route queries to appropriate peers in their acquaintance lists. Therefore, we expect that 6S traffic will become specialized—an effect we call *semantic locality*.¹⁵ Fig. 3 illustrates how semantic

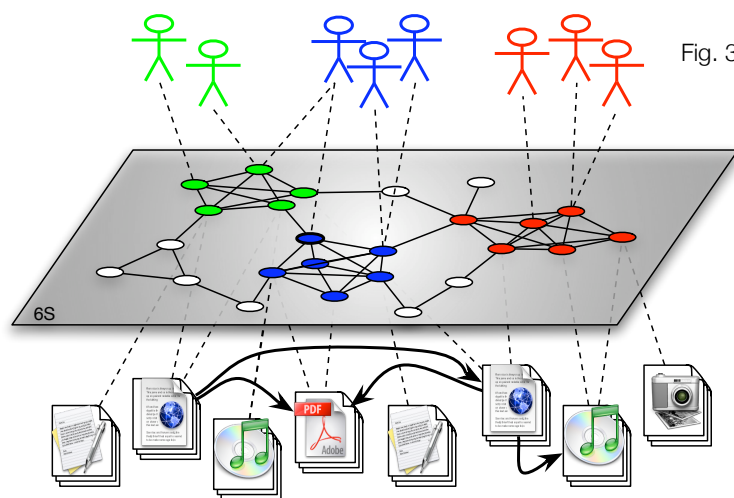


Fig. 3

communities can emerge from 6S clusters. 6S mediates between users and data sources distributed across Web sites and user disks. Data from Web crawls also forms a separate network based on hyperlinks. As peers interact and learn about each other's users and data, the topology of the network adjusts to increase semantic locality and communities emerge. Certain paths will be statistically more used for certain topics than others.

We will study how to discover semantically related neighbors and identify communities and sub-communities (clusters) of interest, even for “new” queries for which appropriate models of peers do not exist locally. This capability will be key for uncovering the underlying semantic locality among peers, so that the bulk of traffic remains localized. Several algorithms typically employed in IR and collaborative filtering may be relevant for this task. For instance, Latent Semantic Analysis to identify the main components of a network, which may be associated with specific semantic interests. Other relevant techniques include vector-based similarity methods, spreading activation,



and link-based community detection algorithms. UIMA can make it easy to explore the use of these techniques. Furthermore, prior applications of UIMA to social networking applications suggest that we can leverage UIMA capabilities to help develop peer community identification algorithms. We need to study how the social network algorithms are affected by the use of query semantics.

Outline plan

The funding of an UIMA Innovation Award would only support one graduate student for one academic year (two semesters). Realistically, this will not suffice to follow all of the research directions outlined in this proposal. The first semester will serve to become acquainted with UIMA and integrating its framework into the 6S prototype, so as to enable extensions of 6S that leverage UIMA components. In this phase we will develop tools to make UIMA accessible to the 6S community, and also consider integrating Apache Lucene/Nutch with UIMA.

The second semester will be devoted to the selective application of UIMA features to 6S. As time allows, we plan to focus on the following:

- annotations of peer collections for refining topical crawls;
- interface for collection management;
- application of text analytics for automatic linking of local resources;
- incorporation of semantic analysis module into query routing;
- analysis of streaming query/response data for peer modeling/learning;
- social networking algorithms.

We realize that this is an ambitious plan but we feel that progress will be possible in a significant portion of the above issues. The main milestone of the project will be a new 6S prototype that will leverage UIMA in various improved functionalities and lend itself to the development of diverse community search services.

Resources

The PI's research group (<http://homer.informatics.indiana.edu/~nan/>) already has an established research program in several aspects of social search; in addition to 6S, another effort (GiveALink.org) relates to building a global semantic similarity network from contributed bookmark. We have already built systems that we continue to extend.

We have ongoing collaborations to many other faculty at IU and with graduate students in Computer Science and Informatics. Most closely connected to the project proposed here are Prof. Luis Rocha, who has an interest in collaborative, community-based digital libraries and recommendation systems; and two PhD students in Computer Science who have worked on the 6S project in the past: Le-Shin Wu and Ruj Akavipat.

We can leverage all of this infrastructure, community, and other sources of funding as we extend the work to look at ways to enhance social search (as described above) with semantic capabilities based on UIMA and at the same time help build the community of people working on and with UIMA, thus enhancing UIMA itself.



References

- ¹ S. Lawrence and C. Giles. Accessibility of information on the Web. *Nature*, 400:107–109, 1999.
- ² J. Cho and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB)*, 2000.
- ³ B. E. Brewington and G. Cybenko. How dynamic is the Web? In *Proc. 9th International World-Wide Web Conference*, 2000.
- ⁴ D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of Web pages. In *Proc. 12th International World Wide Web Conference*, 2003.
- ⁵ A. Ntoulas, J. Cho, and C. Olston. What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.
- ⁶ J. Cho and S. Roy. Impact of search engines on page popularity. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, *Proc. 13th international conference on World Wide Web*, pages 20–29. ACM, 2004.
- ⁷ S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proc. Natl. Acad. Sci. USA*, 103(34):12684–12689, 2006.
- ⁸ M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. In *Proc. 10th International World Wide Web Conference*, 2001.
- ⁹ A. Heydon and M. Najork. Mercator: A scalable, extensible Web crawler. *World Wide Web*, 2(4):219–229, 1999.
- ¹⁰ J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proceedings of the 11th international conference on World Wide Web*, pages 124–135. ACM Press, 2002.
- ¹¹ J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proc. 6th Symposium on Operating System Design and Implementation (OSDI04)*, 2004.
- ¹² S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.
- ¹³ A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Trans. Inter. Tech.*, 1(1):2–43, 2001.
- ¹⁴ R. Akavipat, L.-S. Wu, and F. Menczer. Small world peer networks in distributed Web search. In *Alt. Track Papers and Posters Proc. 13th International World Wide Web Conference*, pages 396–397, 2004.
- ¹⁵ R. Akavipat, L.-S. Wu, A. G. Maguitman, and F. Menczer. Emerging semantic communities in peer web search. In *Proc. ACM CIKM Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR)*, 2006.
- ¹⁶ L.-S. Wu, R. Akavipat, and F. Menczer. Adaptive query routing in peer Web search. In *Proc. 14th International World Wide Web Conference*, pages 1074–1075, 2005.
- ¹⁷ L.-S. Wu, R. Akavipat, and F. Menczer. 6S: Distributing crawling and searching across Web peers. In *Proc. IASTED Int. Conf. on Web Technologies, Applications, and Services (WTAS)*, 2005.
- ¹⁸ M. Singh, B. Yu, and M. Venkatraman. Community-based service location. *Communications of the ACM*, 44(4):49–54, 2000.



-
- ¹⁹ M. Aktas, M. Nacar, and F. Menczer. Personalizing pagerank based on domain profiles. In *Advances in Web Mining and Web Usage Analysis, Proc. 6th SIGKDD Workshop on Web Mining and Web Usage Analysis (WebKDD 2004)*, volume 3932 of LNAI, pages 104–115. Springer, 2006.
- ²⁰ F. Menczer and R. Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the Web. *Machine Learning*, 39(2–3):203–242, 2000.
- ²¹ F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 4(4):378–419, 2004.
- ²² P. Srinivasan, G. Pant, and F. Menczer. A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447, 2005.
- ²³ F. Menczer. Web crawling. In B. Liu, *Web Data Mining: Exploring Hyperlink, Content and Usage Data*, pages 273–322. Springer, 2007.
- ²⁴ G. Pant, S. Bradshaw, and F. Menczer. Search engine – crawler symbiosis. In T. Koch and I. Solvberg, editors, *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Lecture Notes in Computer Science, Vol. 2769, Berlin, 2003. Springer Verlag.



Budget

Prior funding

There is no current funding for the 6S project. The initial steps toward the design and implementation of 6S were supported by the tail end of the PI's NSF Career Award. That grant is now expired.

Funding requested

We seek support for one graduate student to work with the PI on the proposed work. The student will be supported for one academic year according to the budget below. The salary is set by the department, and the benefits are mandated by school policy. If the student has advanced to candidacy the fee remissions may be lower, in which case any leftover funds can be used toward travel to present results at conferences and/or necessary equipment (the PI has a very outdated workstation for the graduate student to work on).

Description	Cost
1 Graduate Student Salary (50% time)	\$13,750.00
Grad Student Fee Remission (mandated; no indirect)	\$7,247.00
Grad Student Health Insurance	\$1,553.00
Total Direct Charges	\$22,550.00
Indirect costs (MTDC, rate 51.5%)	\$7,881.04
Total	\$30,431.04



P.I. Bio Sketch

Filippo Menczer

Associate Professor
Departments of Informatics and Computer Science
School of Informatics
909 Eigenmann Hall
Indiana University, Bloomington

1900 East Tenth Street
Bloomington, IN 52246
Phone: (812) 856-1377
Fax: (812) 856-1995
E-Mail: fil@indiana.edu
<http://informatics.indiana.edu/fil/>

Professional Preparation

University of Rome <i>La Sapienza</i> :	Physics	<i>Laurea</i> 1991
U. California San Diego	Computer Science	M.S. 1994
U. California San Diego	Computer Science & Cognitive Science	Ph.D. 1998

Appointments

University of Iowa	Assistant Professor, Management Sciences	1998-2003
Indiana Univ., Bloomington	Associate Professor, Informatics and Computer Science	2003-present
Indiana Univ., Bloomington	Member, Cognitive Science Program	2003-present

Publications related to proposed project

- R. Akavipat, L.-S. Wu, F. Menczer, A.G. Maguitman: 6S: Emerging Semantic Communities in Peer Web Search. Proc. ACM CIKM Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR), 2006
- P. Srinivasan, F. Menczer, G. Pant: A General Evaluation Framework for Topical Crawlers. *Information Retrieval* 8(3): 417-447, 2005
- L.-S. Wu, R. Akavipat, F. Menczer: 6S: Distributing crawling and searching across Web peers. Proc. IASTED International Conference on Web Technologies, Applications, and Services (WTAS), 2005
- F. Menczer, G. Pant, P. Srinivasan: Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Trans. on Internet Technology* 4(4): 378-419, 2004
- F. Menczer, R.K. Belew: Adaptive Retrieval Agents: Internalizing Local context and Scaling up the Web. *Machine Learning* 39 (2/3): 203-242, 2000



Other significant publications

- S. Fortunato, A. Flammini, F. Menczer, A. Vespignani: Topical interests and the mitigation of search engine bias. *Proc. Natl. Acad. Sci. USA* 103(34): 12684-12689, 2006
- A.G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, A. Vespignani: Algorithmic Computation and Approximation of Semantic Similarity. *World Wide Web Journal* DOI:10.1007/s11280-006-8562-2 (Extended version of WWW2005 paper finalist for best-paper award)
- B. Markines, L. Stoilova, F. Menczer: Social Bookmarks for Collaborative Search and Recommendation. Proc. 21st National Conference on Artificial Intelligence (AAAI), 2006
- F. Menczer: The Evolution of Document Networks. *Proc. Natl. Acad. Sci. USA* 101: 5261-5265, 2004
- F. Menczer: Growing and Navigating the Small World Web by Local Content. *Proc. Natl. Acad. Sci. USA* 99 (22): 14014-14019, 2002

Synergistic activities

- Topical Crawlers: a Java library developed in collaboration with Gautam Pant and Padmini Srinivasan (<http://informatics.indiana.edu/fil/IS/JavaCrawlers/>); a public Java applet for personalized, query-driven, client-based Web crawling by adaptive online agents (<http://myspiders.informatics.indiana.edu/>); and software and data supporting the evaluation of topical crawling algorithms (<http://informatics.indiana.edu/fil/IS/Framework/>). A first prototype of the 6S peer search application is available at <http://homer.informatics.indiana.edu/~nan/6S/>
- GiveALink: site where donated bookmarks are analyzed for building a new generation of social Web mining techniques to search, recommend, personalize and visualize the Web. <http://givealink.org/>
- Program committees: AAAI (2006), SIGIR (2006, 2002), Artificial Life (2006, 2000), WWW (2005, 2004), LinkKDD (2005), SIGIR Workshop on Search and Discovery in Bioinformatics (2004), GECCO (2002), AAMAS (2001), EMO (2001), IEEE ICEC (1996). Associate/guest editor: *IEEE TEC*, *JASIST*. Referee: *PNAS*, *Machine Learning*, *IEEE TKDE*, *JASIST*, *Evol. Comp.*, *IEEE TEC*, *European Physical Journal*, *Artificial Life*, *IJAIT*, *International Journal on Cooperative Information Systems*, *Cognitive Systems Research*, *IEEE TNN*, *IEEE TSMC*, *Network*, *South African Journal of Science*, *BioSystems*
- As Fellow-at-Large of the Santa Fe Institute, Menczer organized a seminar series on “Complex Adaptive Systems” with 18 speakers from industry and academia presenting interdisciplinary research and applications. The series attracted graduate students and created new opportunities for collaborative research. <http://informatics.indiana.edu/fil/CAS/>
- Other open source software in support of instruction and service includes: (1) OAMulator: Web based educational resource supporting the teaching of instruction set architecture, registers, addressing, assembly, programming, and compilers. Perl. Official Resource of the Computer Science Teaching Center (www.CSTC.org). <http://informatics.indiana.edu/fil/OAM/> (2) Recruit: open-source (GNU) Web based system supporting electronic submissions and group-based collaboration for administration of academic job applications. Perl. Listed in GNU Education, OFSET Freeduc, Perl Archive, CGI Resource Index, etc. <http://informatics.indiana.edu/fil/Recruit/>