Registration form (basic details)

1a. Details of the applicant

Title:	Dr.
First Name:	Vera
Initials:	V.
Prefix:	
Surname	Hollink
Male/female	female
Address of correspondence	Centrum Wiskunde en Informatica Science Park 123 1098 XG Amsterdam
Preference for correspondence in English	no
Telephone	+31-(0)20-5924216
Fax:	+31-(0)20-5924199
Website:	http://www.cwi.nl/∼vera

1b. Title of the research proposal

Detecting relations between search queries and search results to improve access to image repositories.

1c. Summary of research proposal

KEYWORDS: web usage mining, query modification, search support

Despite omnipresent web services, such as Flickr and Google Images, finding images remains a difficult task, forcing users to modify their queries several times in order to fulfill their information needs. Our goal is to make search more efficient by developing methods to assist users during query modification.

Existing techniques for studying query modification do not suffice as a base for developing search support. These techniques employ either search log analysis or user studies. Log analysis uses users' queries and the results selected, but not the *meaning* of the queries. As a result, log analysis yields statistics on the overlap in terms between queries, but not their semantic relations. User studies do reveal semantic relations, but are based on limited numbers of users, making statistics about the use of the various relations unreliable.

Our hypothesis is that search support can be improved by combining statistical search log information with semantic information. The first research challenge is to elevate queries from meaningless strings to entities with well defined properties and explicit relations to other entities, by relating the queries to entities from publicly available linked data sources. Statistical methods will be investigated to infer *semantic modification patterns* from frequently occurring relations between queries. The second challenge is to employ these patterns to automatically improve search support.

The techniques developed will be validated in a series of experiments. Offline experiments using the search logs of four large image providers will measure the predictive value of the modification patterns. The impact of the developed search facilities will be evaluated by incorporating them in a live search engine and comparing the search of users who have access to the extra search facilities to users using the original version of the search engine.

1d. Host institution

Centrum Wiskunde en Informatica, Interactive Information Access Group

1e. NWO Division

EW: Exacte wetenschappen

1f. NWO Domain

Beta

2. Research proposal

2a. Scientific/scholarly quality

Research aims

Image search engines allow users to search in large image collections. Despite considerable improvements in recent years in the area of content- and annotation-based image retrieval, finding images remains a challenging task. Users searching for images need to modify their queries several times to fulfill their information needs: on average, users searching for images need 20% more search iterations than users searching for audio and video content [28] and even 80% more iterations than users searching for textual content [18]. Apparently, finding effective queries is a substantial component of image search. The aim of the proposed research is to develop methods to assist users with query modification, making image search as a whole more efficient. We focus on *information gathering* tasks [1, 6, 21, 23, 27]: collecting multiple pieces of information around a single topic. Compared to other types of search tasks, information gathering tasks occur most frequently, take up most of the users' time and involve most query modifications [1, 21].

To support query modification we first need to understand how users modify their queries and how these modifications help to fulfill their information needs. This gives rise to the following research questions:

RQ 1 How to identify meaningful patterns of relations between sequences of queries that are entered in a search session and between queries and selected images?

RQ 2 How to apply query modification patterns to improve the support that image search engines offer to their users?

The hypothesis is that meaningful relations between queries can be identified by combining statistical information gathered from log files with semantic information. Search logs are automatically collected when users use a search engine. Large sources of semantic information have recently become freely available in the form of *linked open data* [2, 3]. Although the potential of these two data sources is widely acknowledged, their combination is largely unexplored.

We propose to match the queries that users enter consecutively during a search session to entities in linked data, thus elevating the queries from meaningless strings to entities of which the semantic relations to other entities are known. Links between the entities will be exploited to find semantic relations between the queries. Comparing statistics about the relations between large numbers of query pairs enables us to identify semantic modification patterns. In the same way, we will use linked data to establish semantic relations between queries that users enter in the beginning of a session and the images that they select later on.

Existing approaches for analyzing query modification behavior cannot identify semantic modification patterns. Automatic analyses of search logs are purely statistical, using only queries that users have entered and clicks they made on search results [16]. Studies employing this type of analysis have classified query modifications based on the overlap between terms [4, 5, 9, 14, 17, 19, 22, 25, 29]. Term-based methods can only classify pairs of queries that have at least one term in common and, therefore, cannot determine relations between queries that are semantically related but share no terms, such as `George Bush` and `Barack Obama`. Furthermore, these methods recognize the modification of `Venus Williams` to `Serena Williams` as a case in which a term has been substituted, but not as relation between two tennis-players or sisters.

In the field of human-computer interaction, a popular means for researching search behavior is to carry out in-depth studies of the search behavior of a small number of users (e.g. [8, 15]). The (modification) behavior of the users is examined in laboratory experiments or field studies and the users' motivations for the various search steps are revealed through diaries or interviews [20]. In contrast to log analysis, this type of research can reveal semantic relations between queries. However, it is necessarily restricted to a small number of users, which makes statistics about the use of the various types of relations unreliable. Moreover, user studies are costly and time-consuming. Our approach combines the strengths of the two existing approaches, automatically identifying semantic relations and employing log statistics to determine the importance of the various types of relations.

Various authors have advocated the use of term-based modification patterns for enhancing search support, in particular through suggestions for follow-up queries [14, 17, 25], but so far this has not been realized. We will take research into query modification a step further by investigating how modification patterns can be applied to improve support during various phases of information gathering tasks and evaluating the support in user studies:

- In the first phase of the search the user (re)formulates a query, translating the information needs into a textual string. We aim to reduce the number of search iterations by offering suggestions for follow-up queries. Preliminary results have demonstrated that semantic modification patterns can be successfully applied for query suggestion [12]. In contrast to existing suggestion methods based on log data, the pattern-based approach can also generate suggestions for first-time queries. Moreover, the patterns provide explanations of the relations between the query and the suggestions, which in other domains has shown to increase the users' acceptance of system-generated recommendations [7].
- During the second phase the search engine retrieves a set of images that match the user's query. We will investigate under what conditions relations between queries and selected results are usable for predicting from a user's initial queries the images that s/he will select later in the session. The predicted images will be added to the results of the initial queries with the aim of increasing recall and improving the ranking of the search results.
- The final phase involves presenting the search results to the user. We will use modification patterns to explain how the search results retrieved relate to the user's query and to group results by the relations they bear to the query. We expect that this will simplify the selection of relevant search results for the users, shortening the time they need to assess whether results are relevant and reducing the number of clicks on irrelevant results.
- We will personalize the search support for all three search phases by adapting it to the characteristics of specific user groups. Clustering techniques, such as those in [30], will be used to find groups of users with similar modifications patterns. This enables us to base the search facilities on patterns of relevant user clusters instead of patterns of the whole user population, which can reduce the users' search efforts even further.

Approach

During the first six months of the project we will develop the core method for finding semantic relations between consecutive search queries. This method matches queries to entities in linked data by examining the similarity between the queries and the labels of the entities in the linked data. To find the relation between the queries, it searches the shortest series of links that connect the entities. Graph search techniques will be employed to reduce the complexity of this shortest path problem. For example we match the query `Andre Agassi` to the concept `http://dbpedia.org/resource/Andre_Agassi` and the query `Boris Becker` to `http://dbpedia.org/resource/Boris_Becker`. Following the relations in the linked data, we can determine relations between the two concepts, such as that both are men and both are tennis players. This process is illustrated in Figure 1. Often multiple relations exist between two entities. Relations that are likely to be important for users are those that occur significantly more frequently between consecutive query pairs than between query pairs from different search sessions, as we will determine through statistical analysis. Semantic modification patterns will be identified by extracting important relations that occur in many search sessions. In preliminary experiments this method revealed many interesting patterns [13].

The linked data that we will use is part of the publicly available data cloud collected in the Linking Open Data Project [24]. According to the latest estimates this data cloud connects over a hundred data sources, in total consisting of 4.7 billion RDF triples, where each triple defines a property of an entity or a relation between two entities [3]. As the data cloud continues to grow, the methods that we develop will be able to handle more and more different queries.

Previous studies on query modifications were limited to the search logs of one search engine [4, 5, 9, 14, 17, 19, 22, 25, 29]. We will use the search log data of four image search engines: a Dutch and a Belgian

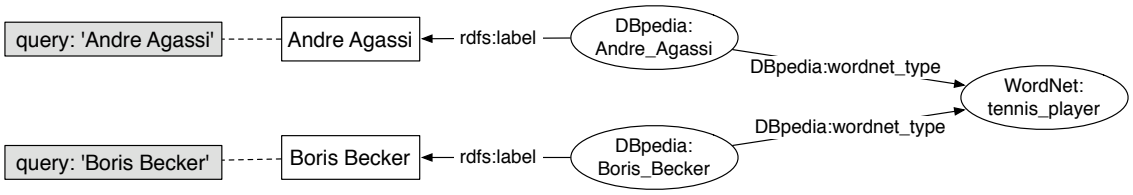

Figure 1: Example application of our method for finding semantic relations between queries: a relation between queries `Andre Agassi` and `Boris Becker` is that they both match DBpedia entities that are of WordNet type `tennis_player`.

vendor of historic and news photos, a web site of a Dutch art museum and a portal providing access to many cultural collections in Europe. These institutes have made their log data available to the research group as part of longer term collaborations. Having search logs from various partners provides us with the unique opportunity to compare modification patterns in different data sets and determine which patterns are valid across different types of image search and which differ among search engines and user populations.

In the second period of six months we will investigate methods to apply semantic modification patterns for query suggestion. We will evaluate the various approaches using the log data of the four image search engines. In this stage the suggestions will be evaluated offline by measuring how often the next query of a user in the log file is among the suggested queries [10].

The second year will be devoted to identifying and exploiting relations between queries and selected images. We will use terms from the image annotations to match images to concepts in linked data sources. Patterns of these relations will be used to predict which images will be relevant for later queries and to add these to the result lists. The effects of the added images on precision and recall will again be assessed offline.

In the third year we will focus on employing the patterns to improve result presentation. Moreover, all search facilities developed so far will be subjected to a user study. For this, we will make use of an existing search engine for cultural heritage content, which is used by a considerable number of real users. We will incorporate the various facilities into the search engine and offer them as extra support to a random selection of the search engine's users. Differences between the search behavior of users with and without access to the various facilities will allow us to accurately measure the added value of each type of search support.

The main topic of the final year will be search personalization. We will investigate methods to recognize user groups with different modification patterns. The search facilities that proved most successful in the user experiment will be augmented with personalization based on these user groups. The personalization will be assessed in a second user experiment.

In the end we expect that our methods will improve the efficiency and efficacy of image search by significantly decreasing the number of iterations users need to fulfill their information needs and increasing the number of relevant images found.

Time line

Months	Topic	Deliverables
1-6	method for identifying semantic modification patterns	conference paper
7-12	employing patterns for query suggestion	journal paper
13-18	method for identifying relations between queries and search results	conference paper
19-24	employing patterns to improve recall	journal paper
25-30	employing patterns to improve search result presentation	conference paper
31-36	large scale user evaluation	conference and journal paper
37-48	personalization	journal paper

One month equals 0.75 person months of work.

Research environment

The Interactive Information Access Group of the Centrum Wiskunde en Informatica consists of leading researchers including Dr. Van Ossenbruggen, Prof.Dr.Ir. De Vries, and Prof.Dr. Hardman. The group combines experience on (image) information retrieval with research on using semantics for information disclosure. In the context of various international projects and networks, such as Vitalas, EuropeanaConnect, and PetaMedia, the research group has established several lasting collaborations with organizations that maintain image search engines. These collaborations enable us to work with the search log data of large numbers of real users. Moreover, contacts within these institutes provide a platform to discuss the demands of real image providers and for getting feedback on our solutions. On the topic of using semantics in search the group closely collaborates with the VU Web and Media Group. This collaboration has yielded an award-winning prototype [26], which will serve as a platform on which our results will be evaluated.

2b. Research impact

As increasing numbers of images are becoming available in digital image repositories, methods for efficient disclosure of the repositories are urgently needed. The proposed research contributes to the disclosure of digital image collections by making use of the feedback that is implicitly provided when people are using the system. In order to carry out our studies we need data sets and log files, which are available for cultural heritage archives and commercial image providers. However, the proposed methods are by no means restricted to these domains: semantic modification patterns have the potential to enhance search in a wide variety of visual collections, in particular where domain knowledge is available in the form of thesauri.

Cultural heritage archives

At present almost all cultural heritage institutes (museums, archives, libraries) have digitized or are in the process of digitizing the contents of their paper archives. The images are often annotated with terms from domain-specific thesauri. In various large scale projects collections and thesauri of different archives are being connected to each other and to general purpose ontologies, allowing users to search through even larger image collections. The collections are queried by both amateurs and professionals users, such as art historians and museum curators. Queries of professional users tend to be broad and complex, demanding more advanced search techniques than the simple keyword matching offered by most existing search facilities [1, 11].

Our research provides the technologies needed to develop search facilities that help users to answer complex questions faster and more easily. Suggestions for related concepts help users to broaden their search. Explicit explanations for the relations between queries and search results allow users to recognize the relevance of non-obvious results. The specialized thesauri available in the domain, linked with general purpose lexical resources such as WordNet, enable us to provide support for both queries with domain-specific terms and queries with layman terminology.

We apply our methods to the search logs of the search engines of two cultural heritage archives that have made their log data available to the research group: a Dutch art museum and a portal providing access to many European cultural collections. If the results of the evaluation of the created search support is positive, we expect, based on our collaborations with the institutes, that the institutes will incorporate the support in their search engines.

Commercial image providers

Commercial image providers offer vast collections of stock images related to a multitude of topics, such as historical events, news and sports. New content is constantly added to the collections and new topics are emerging all the time. Users have diverse backgrounds and purchase images for a variety of applications. As a result, the system is constantly confronted with users with new information needs, formulating queries that not have been entered before and searching for newly added images.

The dynamic nature of the collection makes the domain unsuitable for search facilities that rely on statistics about the use of individual images and queries. The semantic modification patterns that we identify describe the search behavior of the users on a higher level of abstraction, which makes search

facilities based on these patterns robust against changes in content and queries. Personalization techniques that we develop enable us to deal with the diversity in the user population.

Search logs of a vendor of historic photos and a vendor of news photos will be used to test our techniques in the domain of commercial image providers. To these companies the developed techniques are of direct commercial value, as they enable their customers to search the collections more effectively.

Other potential application areas

Medical image collections contain a wealth of information relevant for health-care professionals as well as patients. The collections are growing rapidly due to images collected through daily clinical routines, which heightens the need for effective search support. The medical domain is characterized by a large number of synonyms and a large diversity between terms used by laypeople and medical specialists. As image collections are usually annotated with terms from only one vocabulary, queries using other terminologies often yield no results. Another difficulty is that a dense network of relations exists between medical concepts, many of which are not immediately clear to non-experts. Not showing images annotated with related concepts deprives the user of many potentially useful results. On the other hand, showing all related images makes it difficult for users to understand why the images presented are relevant. Medical ontologies combined with automatically collected log data from search engines for medical images enable the implementation of search facilities based on semantic modification patterns. These facilities provide adequate solutions for the search problems in this domain. Through medical ontologies we can match laypeople's terms with highly specialized terms, enabling us to generate query suggestions that help users to find terms that match the vocabulary of the collection. Furthermore, our methods provide explanations for relations between the user's query and the images that are retrieved.

The rise of digital cameras and affordable storage solutions has dramatically increased the amount of image material in **personal picture collections**. The images in these collections are usually sparsely annotated and the annotations that exist use inconsistent terminologies. One collection often contains many images from the same event or on the same topic. Users searching in their personal collections often want all images on one topic, demanding high recall. This is difficult with the existing search engines as they are mostly focused on precision. The methods that we develop in this project can enhance recall as they facilitate finding images annotated with a terminology different from the user's query and broadening of search to closely related topics.

The techniques that we develop can also be generalized to other media, such as **video**. Also in video collections, searchers experience considerable difficulties formulating adequate queries and need to try various queries before they find what they need [18]. We provide methods to analyze these query modifications and use this information to assist users during their search.

2c. Number of words used:

Section 2a: 1998 words (maximum number of 2000 words)
Section 2b: 993 words (maximum number of 1000 words)

2d. Any other important remarks with regard to this application

-

2e. Literature references

[1] Alia Amin, Jacco van Ossenbruggen, Lynda Hardman, and Annelies van Nispen. Understanding cultural heritage experts' information seeking needs. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries, Pittsburgh PA, USA*, pages 39–47, 2008.

[2] Tim Berners-Lee. Linked data: Design issues. http://www.w3.org/DesignIssues/LinkedData.html, 2006. last accessed November 5, 2009.

[3] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems, Special Issue on Linked Data*, in press.

[4] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. From 'dango' to 'japanese cakes': query reformulation models and patterns. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy*, pages 183–190, 2009.

[5] Peter Bruza and Simon Dennis. Query reformulation on the internet: empirical data and the hyperindex search engine. In *Proceedings of the RIAO'97 Conference on Computer-Assisted Searching on the Internet, Montreal, Canada*, pages 488–499, 1997.

[6] Chun Wei Choo, Brian Detlor, and Don Turnbull. Working the web: an empirical model of web use. In *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 7*, 2000.

[7] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455–496, 2008.

[8] Efthimis N. Efthimiadis. Interactive query expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003, 2000.

[9] Daqing He, Ayse Göker, and David J. Harper. Combining evidence for automatic web session identification. *Information Processing and Management*, 38(5):727–742, 2002.

[10] Qi He, Daxin Jiang, Zhen Liao, Steven C. H. Hoi, Kuiyu Chang, Ee-Peng Lim, and Hang Li. Web query recommendation via sequential query prediction. In *Proceedings of the 25th IEEE International Conference on Data Engineering, Shanghai, China*, pages 1443–1454, 2009.

[11] Michiel Hildebrand, Jacco Van Ossenbruggen, Lynda Hardman, and Jan Wielemaker. Towards reasoning patterns for semantic search: a case study in cultural heritage. *Forthcoming*.

[12] Vera Hollink, Theodora Tsikrika, and Arjen de Vries. The semantics of query modification. *Submitted for publication*.

[13] Vera Hollink, Theodora Tsikrika, and Arjen de Vries. Semantic vs term-based query modification analysis. In *Proceedings of Tenth Dutch-Belgian Information Retrieval Workshop, Nijmegen, the Netherlands*, 2010.

[14] Jeff Huang and Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China*, pages 77–86, 2009.

[15] Michael Huggett and Joel Lanir. Static reformulation: a user study of static hypertext for query-based reformulation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries, Vancouver, BC, Canada*, pages 319–328, 2007.

[16] Bernard J. Jansen. Search log analysis: what it is, what's been done, how to do it. *Library and Information Science Research*, 28(3):407–432, 2006.

[17] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Patterns of query reformulation during web searching. *Journal of the Americal Society for Information Science and Technology*, 60(7):1358–1371, 2009.

[18] Bernard J. Jansen, Amanda Spink, and Jan O. Pedersen. The effect of specialized multimedia collections on web searching. *Journal of Web Engeneering*, 3(3-4):182–199, 2004.

[19] Corinne Jörgensen and Peter Jörgensen. Image querying by image professionals. *Journal of the Americal Society for Information Science and Technology*, 56(12):1346–1359, 2005.

[20] Melanie Kellar, Kirstie Hawkey, Kori M. Inkpen, and Carolyn Watters. Challenges of capturing natural web-based user behaviours. *International Journal of Human Computer Interaction*, 24(4):385—409, 2008.

[21] Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the Americal Society for Information Science and Technology*, 58(7):999–1018, 2007.

[22] Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. In *Proceedings of the Seventh International Conference on User Modeling, Banff, Canada*, pages 119–128, 1999.

[23] Julie B. Morrison, Peter Pirolli, and Stuart K. Card. A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA*, pages 163–164, 2001.

[24] Linking Open Data Project. Linked open data W3C SWEO community project, 2009. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData, Last accessed November 19, 2009.

[25] Soo Young Rieh and Hong Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management*, 42(3):751–768, 2006.

[26] Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Osenbruggen, Anna Tordai, Jan Wielemaker, and Bob Wielinga. Semantic annotation and search of cultural-heritage collections: the multimedian e-culture demonstrator. *Journal of Web Semantics*, 6(4):243–249, 2008.

[27] Abigail J. Sellen, Rachel Murphy, and Kate L. Shaw. How knowledge workers use the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Minneapolis, MN, USA*, pages 227–234, 2002.

[28] Dian Tjondronegoro, Amanda Spink, and Bernard J. Jansen. A study and comparison of multimedia web searching: 1997-2006. *Journal of the American Society for Information Science and Technology*, 60(9):1756–1768, 2009.

[29] Martin Whittle, Barry Eaglestone, Nigel Ford, Valerie J. Gillet, and Andrew Madden. Data mining of search engine logs. *Journal of the American Society for Information Science and Technology*, 58(14):2382–2400, 2007.

[30] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the Tenth International Conference on Information and Knowledge Management, Atlanta, GA*, pages 25–32, 2001.

3. Cost estimates

3a. Budget

Staff costs: (in k€ incl. surcharge)							
	FTE	Nr. of months	Year 1	Year 2	Year 3	Year 4	TOTAL
Applicant Support staff	0.75	48	47.1	48.7	50.2	51.7	197.7
Staff costs: (in k€ incl. surcharge)							
			Year 1	Year 2	Year 3	Year 4	TOTAL
Equipment			2.0				2.0
Travel and subsistence			6.0	6.0	6.0	6.0	24.0
TOTAL			55.1	54.7	56.2	57.7	223.7

3b. Indicate the time (percentage of fte) you will spend on the research

0.75 fte (= 100% of employment)

3c. Intended starting date

November 1, 2010

3d. Have you requested any additional grants for this project either from NWO or from any other institution?

no

3e. Has the same idea been submitted elsewhere

no

4. Curriculum vitae

4a. Personal details

Title(s), initial(s), First name, surname:	Dr., V., Vera, Hollink
Male/female:	female
Date and place of birth:	November 13, 1980, Haarlem, The Netherlands
Nationality:	Dutch
Birth country of parents:	The Netherlands

4b. Master's ('doctoraal')

University/College of Higher Education:	Universiteit van Amsterdam
Date (dd/mm/yy):	13/12/02
Study/main subject:	Artificial intelligence with specialization language and logic

4c. Doctorate

University/College of higher education:	Universiteit van Amsterdam
Date (dd/mm/yy):	31/01/08
Supervisor('Promotor'):	Prof.Dr. B. J. Wielinga
Title of Thesis:	Optimizing hierarchical menus: A usage-based approach

4d. Use of extension clause

no

4e. Current employment

Postdoctoral researcher at the Interactive Information Access Group of the Centrum Wiskunde en Informatica. Fixed term contract: 3 years, 0.7 fte.

4f. Work experience since graduating

March 2009 - present	Postdoctoral researcher Centrum Wiskunde en Informatica 0.7 fte, fixed term
September 2007 - February 2009	Postdoctoral researcher University of Amsterdam 0.9 fte / 0.7 fte, fixed term
February 2003 - August 2007	Doctorate research University of Amsterdam 1 fte / 0.9 fte, fixed term

4g. Man-years of research

Period	Position	Man years/months
February 2003 - January 2005	1 fte	2 years
February 2005 - August 2007	0.9 fte minus 4 months sick-leave	2 years
September 2007 - June 2008	0.9 fte	9 months
July-October 2008	maternity leave	
November 2008 - December 2009	0.7 fte	9 months
Total		5 years and 6 months

Since November 2008 I have worked 0.7 fte to combine work with care for my child.

4h. Brief summary of research over the last five years

My research aims at enhancing the efficiency and efficacy of information search. It has always been strongly interdisciplinary, combining a human-computer interaction perspective with techniques from information retrieval and machine learning. Although my ultimate goal is to automatically build search support tools, understanding how users interact with information systems is a large aspect of my work. This perspective ensures that all developed methods for supporting information search are firmly grounded in realistic models of user behavior. Among the most important outcomes of my research are several novel techniques for analyzing user behavior recorded via interaction logs.

My Ph.D. research focused on web browsing. I developed techniques for analyzing browsing behavior, especially in hierarchical link structures. In addition, I presented methods to automatically improve link structures based on the outcomes of the analyses. The effectiveness of the techniques was proven in a series of user experiments. Moreover, for this project I collaborated with researchers from TNO Quality of Life, who incorporated part of the developed technology in a professional health care web site. In my first postdoc project I moved from support for web users to supporting web site authoring. I developed a system to automatically compare contents of web sites, which combined machine learning with information retrieval techniques. In my current postdoc position I extend my Ph.D. work from browsing to keyword search. I analyze how users translate their information needs into search queries and use this information to create tools that support users with the search.

4i. International activities

Visiting researcher:

- School of Computing and Mathematics, University of Ulster, January-June 2004. Conducted as part of my Ph.D. research under the supervision of Dr. S.S. Anand on the subject of predicting web navigation behavior.

Program Committee member:

- The First International Conference on Adaptive and Self-adaptive Systems and Applications, 2009.

Reviewer International Journals:

- User Modeling and User Adapted Interaction
- World Wide Web Journal

4j. Other academic activities

Undergraduate supervisor:

- 3 M.Sc theses 2005-2006
- 3 B.Sc theses 2003-2008.

Lecturer:

- Graduate course on Internet Information, together with Dr. C. Monz, University of Amsterdam, 2010
- Graduate course on Machine Learning, together with Dr. M. van Someren, University of Amsterdam, 2003 and 2004.

4k. Scholarships, grants and prizes

Marie-Curie scholarship within PERSONET to visit the School of Computing and Mathematics, University of Ulster, 2004.

5. Publications

Theses

- V. Hollink. Optimizing Hierarchical Menus: A Usage-based Approach. *Ph.D. thesis, University of Amsterdam, Amsterdam, The Netherlands*, 2008.
- V. Hollink. Algoritmes voor Pronoun Resolutie: Een Vergelijkende Studie en een Aanzet tot een Nieuwe Statistische Methode. *Master thesis, University of Amsterdam, Amsterdam, The Netherlands*, 2002.

International refereed journals

- V. Hollink, M. van Someren and B. J. Wielinga. A Semi-Automatic Usage-Based Method for Improving Hyperlink Descriptions in Menus. *International Journal of Human-Computer Studies* 67, pages 366-381 , 2009. (ISI impact factor 1.769)
- V. Hollink, M. van Someren and B. Wielinga . Navigation behavior models for link structure optimization. *User Modeling and User-Adapted Interaction* 17 (4), pages 339-377, 2007. (ISI impact factor 1.483)
- V. Hollink, M. van Someren and B. Wielinga . Discovering stages in web navigation for problem-oriented navigation support. *User Modeling and User-Adapted Interaction, special issue on Statistical and Probabilistic Methods for User Modeling* 17 (1-2), pages 183-214, 2007. (ISI impact factor 1.483)
- V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual Document Retrieval for European Languages. *Information Retrieval* 7(1-2), pages 33-52, 2004. , (ISI impact factor 2004 1.396)

International refereed conferences

- V. Hollink, V. de Boer, M.W. van Someren. SiteGuide: A Tool for Web Site Authoring Support. To appear in: *J. Cordeiro and J. Filipe (eds.): WEBIST 2009 Revised Best Papers, Lecture Notes in Business Information Processing*, Springer.
- V. Hollink, M.W. van Someren, V. de Boer. Clustering Objects from Multiple Collections. *Proceedings of 32nd Annual Conference on Artificial Intelligence, Paderborn, Germany*, 2009.
- V. de Boer, V. Hollink and M.W. van Someren. Automatic Web Site Authoring with SiteGuide. *Proceedings of Intelligent Information Systems, Krakow, Poland*, 2009.
- V. Hollink, M. van Someren and V. de Boer. SiteGuide: an Example-based Approach to Web Site Development Assistance. *Proceedings of the Fifth International Conference on Web Information Systems and Technologies, Lisboa, Portugal*, 2009.
- V. Hollink, M. van Someren and S. ten Hagen. Discovering Stages in Web Navigation. *Proceedings of the 10th International Conference on User Modeling, Edinburgh, UK*, 2005.
- S. ten Hagen, M. van Someren, and V. Hollink. Exploration/Exploitation in Adaptive Recommender Systems. *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and their Implementations in Smart Adaptive Systems, Oulu, Finland*, 2003.

Local refereed conferences

- V. Hollink, T. Tsikrika and A. P. de Vries. Semantic vs Term-based Query Modification Analysis. Optimal Link Categorization for Minimal Retrieval Effort. To appear in: *Proceedings of 10th Dutch-Belgian Information Retrieval Workshop, Delft, The Netherlands*, 2010.

- V. de Boer, V. Hollink and M.W. van Someren. Automatic Web Site Authoring with SiteGuide. *Proceedings of the 21st Belgian-Netherlands Conference on Artificial Intelligence, Eindhoven, the Netherlands*, 2009.
- J. Mostert, V. Hollink. Effects of Goal-Oriented Search Suggestions. *Proceedings of the 20th Belgian-Netherlands Conference on Artificial Intelligence, Boekelo, The Netherlands*, 2008.
- V. Hollink, M. van Someren. Web usage mining for the classification of link anchors. *Proceedings of the sixteenth Benelearn Conference, Amsterdam, The Netherlands*, 2007.
- V. Hollink and M. van Someren. Optimal Link Categorization for Minimal Retrieval Effort. *Proceedings of 6th Dutch-Belgian Information Retrieval Workshop, Delft, The Netherlands*, 2006.

International refereed workshops

- V. Hollink, M. van Someren and V. de Boer. Capturing the Needs of Amateur Web Designers by Means of Examples. *Proceedings of the Annual Workshop of the SIG Adaptivity and User Modeling in Interactive Systems 2008, Würzburg, Germany*, 2008.
- V. Hollink, M. van Someren, S. ten Hagen, M.J.C. Hilgersom, J.M. Rövekamp. The SeniorGezond Recommender: Exploration Put into Practice. *Proceedings of the AAAI Workshop on Intelligent Techniques for Web Personalization, Vancouver, Canada*, 2007.
- V. Hollink, M. van Someren and B. Wielinga . Using log data to detect weak hyperlink descriptions. *Proceedings of the UM'07 Workshop on Data Mining for User Modeling, Corfu, Greece*, 2007.
- V. Hollink and M. van Someren. Validating Navigation Time Prediction Models for Menu Optimization. *Proceedings of the 14th Workshop on Adaptivity and User Modeling in Interactive Systems, Hildesheim, Germany*, 2006.
- V. Hollink, M. van Someren, S. ten Hagen and B. Wielinga. Recommending Informative Links. *Proceedings of the IJCAI-05 Workshop on Intelligent Techniques for Web Personalization, Edinburgh, UK*, 2005.
- V. Hollink, M. van Someren and S. ten Hagen. Web Site Adaptation: Recommendation and Automatic Generation of Navigation Menus. *Proceedings of the Annual Workshop of the SIG Adaptivity and User Modeling in Interactive Systems 2004, Berlin, Germany*, 2004.
- M. van Someren, S. ten Hagen, and V. Hollink. Greedy Recommending is not Always Optimal. *B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, and G. Stumme (eds.): Web Mining: From Web to Semantic Web, Lecture Notes in Artificial Intelligence* 3209, pages 148-163, Springer, 2004.
- M. van Someren, S. ten Hagen, and V. Hollink. Greedy Recommending is not Always Optimal. *Proceedings of the 1st European Web Mining Forum, Cavtat-Dubrovnik, Croatia*, 2003.

Statements by the applicant

I have completed this form truthfully.

Name:	Vera Hollink
Place:	Amsterdam
Date:	January 7, 2010