Future Internet **2010**, *2*, 533-558; doi:10.3390/fi2040533

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Ontology-Based Information Behaviour to Improve Web Search

Silvia Calegari * **and Gabriella Pasi**

Department of Informatics, Systems and Communication-DISCo, University of Milano-Bicocca, V.le Sarca 336-14, 20126, Milano, Italy; E-Mail: pasi@disco.unimib.it

* Author to whom correspondence should be addressed; E-Mail: calegari@disco.unimib.it; Tel.: +39 0264487917; Fax: +39 0264487880.

Received: 24 September 2010; in revised form: 11 October 2010 / Accepted: 13 October 2010 / Published: 18 October 2010

Abstract: Web Search Engines provide a huge number of answers in response to a user query, many of which are not relevant, whereas some of the most relevant ones may not be found. In the literature several approaches have been proposed in order to help a user to find the information relevant to his/her real needs on the Web. To achieve this goal the individual Information Behavior can been analyzed to 'keep' track of the user's interests. *Keeping* information is a type of Information Behavior, and in several works researchers have referred to it as the study on what people do during a search on the Web. Generally, the user's actions (e.g., how the user moves from one Web page to another, or her/his download of a document, *etc.*) are recorded in Web logs. This paper reports on research activities which aim to exploit the information extracted from Web logs (or query logs) in personalized user ontologies, with the objective to support the user in the process of discovering Web information relevant to her/his information needs. Personalized ontologies are used to improve the quality of Web search by applying two main techniques: query reformulation and re-ranking of query evaluation results. In this paper we analyze various methodologies presented in the literature aimed at using personalized ontologies, defined on the basis of the observation of Information Behaviour to help the user in finding relevant information.

Keywords: keeping information; ontologies; query reformulation; re-ranking

---

1. Introduction

In the last few years there has been an exponential growth of the information available on the Web. The difficulty of locating information relevant to the user's needs is becoming a 'hot' and crucial problem, and a great deal of research is addressing this topic [1,2]. People mainly use Web search engines as an access point to relevant information on the Internet. Unfortunately, a search engine produces several results in response to a user query, with the consequence that the truly relevant results are not always retrieved. In the literature, in order to face this unsatisfying situation, one of the suggested solutions is related to the contextualization of a user's Web search. The objective is to personalize the behaviour of Web search engines [2]. Indeed, the results produced by a search engine for a given query are the same, independent of the user and the context in which the user made the request. The main idea is to understand what the context is surrounding a Web-query during a user session: in the literature several proposals have been defined for this [3]. Possible solutions are related to *query-specific contexts*, and they include: (i) *context around query* and (ii) *context within query* [4]. The former specifies the environment of a query, such as the domain of interest, while the latter refers to context words within the query [4].

The problem of personalization has also been faced in several studies of Information Behaviour. In Wilson's model of information behaviour [5] the context of the information need is based on the personal context of a user where information seeking and searching are jointly considered at different stages of a search session. In particular, the personal context of a user is defined by collecting the information *acquired and organized over time and in response to a range of stimuli* [6]. On the Web, the search context may be defined by considering the information collected over time after several user's queries. This means analyzing the behaviour of the individuals: how a user moves from a Web page to another, what the actions performed on a Web page are, what information is stored in the user's workstation, *etc.* The user's actions are organized into Web logs and may be re-used in future user queries in order to help the user locate relevant information during a Web search. In this scenario, the assumption is that a user only keeps track of the useful information while discarding the rest. In [7] an analysis of some people-centred information applications developed to support the user during a Web search is reported.

In this paper, an overview of the use of ontologies to improve Web search is presented. In particular, we consider the approaches which make use of ontologies to formally represent the information behaviour of the individuals during search sessions on the Web. Ontologies give organization to the information, and they allow associations between data to be expressed. In this context two types of ontologies are involved: (1) personalized ontologies, and (2) general purpose ontologies (e.g., the ODP [8] ontology). In both cases, the ontology is used as a semantic support to capture the user's intents in a Web search.

A personalized ontology can be semi-automatically defined by considering: (1) the terms used in the user's queries, (2) the user's activities and/or 93) the user's preferred documents. These different strategies can be jointly used in order to obtain a personalized ontology. In the first case the words written in the user's queries are stored over time into query log files (past queries) to discover associations between terms [9,10]. In this field, one of the pioneer works is the one proposed by Silverstein *et al.* [11]. In the second case, the user's actions during a Web session are analyzed and then stored into Web log files. Generally, a Web log file includes information like a unique identifier for

the user or a session, a query string, a timestamp, the results' page number, the URLs clicked for each query and actions such as saving, printing, copy, *etc.* [12]. In the third case, an ontology is defined by extracting the relevant information from the user's preferred documents. These documents are stored on the user's workstation during previous Web sessions [13]. A personalized ontology is then obtained by keeping the user's useful information after a user search session according to these three possibilities.

In this work we present a review of some works proposed in the literature on the use of a personalized ontology to help the user find the information relevant to his/her needs. In particular, we describe how an ad-hoc ontology can be used to improve the quality of a user's Web search by facing two typical Information Retrieval problems: query reformulation and result re-ranking. In the first case some useful concepts defined in the ontology are used to expand the user's query in order to better locate relevant information; in the second case, the ontology is used to re-rank the results produced by a search engine according to the user's interests.

This paper is organized as follows. In Section 2 the use of ontologies to model Information Behaviour is presented. Section 3 and Section 4 report a review of the use of an ontology to personalize the query expansion and re-ranking strategies, respectively. In Section 5 some conclusions are stated.

2. Use of Ontologies to Capture the Information Behaviour

Today's search engines constitute an important search tool on the Web for many people; although search engines usually provide a huge numbers of answers, many of which are not relevant, some of the more interesting answers are not found. One of the reasons for this situation is that existing Web search engines are not able to *interpret* the user's context (at least to some extent). To improve this situation it would be useful to give the systems the possibility to better understand the meaning of a user's query in order to produce better answers. The goal, in this case is to reduce the communication gap between humans and machines: *"... interaction between people and computers requires essentially the same interpretative work that characterizes interaction between people"* [14].

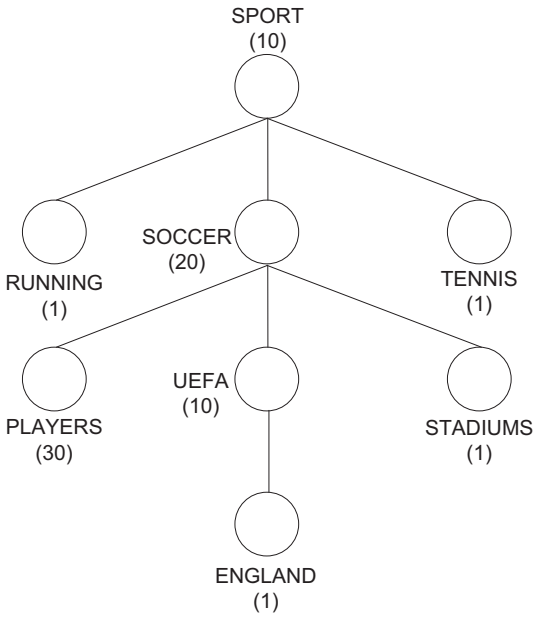
Ontologies have been used as a possible approach towards the above research direction; in fact, as they allow to represent the knowledge of a specific domain, ontologies can be used to help people and machines to communicate concisely by facilitating the information exchange based on semantics rather than just on syntax. An ontology is made up of concepts, instances, functions, relationships and axioms to provide a common understanding of the real world [15–18]. With the support of ontologies, users and systems can communicate with each other through an easy information integration [19].

Ontologies can be employed not only to reduce the communication gap between humans and machines but also to formally represent the user's interests and context. The definition of a personalized ontology is aimed at representing the user's personal interests. Since generally a person has several interests, user profiles should be defined to represent the various user interests. For example, in [20] a user profile is proposed to represent the distinct interests related to a user. The use of an ontological language to formally define a user profile can offer an effective solution to this problem. For this reason, in several works, a personalized ontology is viewed like a user profile, named *ontological user profile* [21]. To understand the user intent of a Web search could in this context mean identifying the user's interest related to the query in the user ontology.

Generally, an ontology can be defined either manually by experts of the considered domain or in a semi-automatic way by analyzing a specific source (domain-dependent) of information. In the literature two main approaches have been applied to define an ontological user profile: (1) by the use of general purpose ontologies (e.g., ODP), and (2) by personalized ontologies. In both cases, the ontology is used as a semantic support to identify user intentions.

In [21] an example of a personalized ontology obtained from the ODP is presented. Figure 1 depicts a graphical representation of an ontological user profile where just a *portion* of the knowledge is represented in the ODP. In particular, only the categories of user interest are automatically selected from the ODP after the analysis of the user's actions during a search session, and an importance score is associated with each concept (for more details see Section 3). In this quite simple representation the categories are linked through the taxonomic relation IS-A.

Figure 1. A simple example of an ontological user profile.



In some Information Behaviour works, a personalized ontology is employed to provide a semantic organization to the information selected by the users within a search session. In this situation a personalized ontology is (semi-)automatically built after the analysis, for example, of Web log files. In [22] a domain model based on an ontology, named *Behavior Knowledge Base*, is defined by storing the information about all the significant actions performed by the user on the Web. Examples of significant user's actions are: accessing a Web page, downloading the preferred information, *etc.*

Once a personalized ontology has been defined, it is possible to use it in order to improve the user's search. The approaches proposed in the literature address the definition of methodologies to identify the useful information in the ontology to improve a search task [6,23]. Two main approaches have been investigated: query expansion, and re-ranking of the results produced by a search engine in response to a user's query. In query expansion a personal ontology can be used to automatically expand the user's search, in order to contextualize it. Whereas in the re-ranking phase, a personal ontology can be used to re-rank the search result based on the conceptual schema provided by the user ontology structure. In the literature, hybrid approaches are also proposed. In [24] a survey of methodologies of

concept-based query expansion for re-ranking multimedia documents is reported. The query expansion process (especially for short query) can improve recall, whereas the re-ranking phase can improve precision; thus techniques based on hybrid methodologies are preferred [24]. In Section 3 an overview of some works presented in the literature concerning the use of a personalized ontology aimed at query expansion is reviewed; in Section 4 an overview of some strategies defined in the literature employed for re-ranking search results according to the user interests is presented.

The Open Directory Project As previously outlined, an ontological user profile is essentially a portion of a reference knowledge, where each concept refers to the user's perceived interest. In this short review of ontology-based approaches to personalization, the Open Directory Project (ODP) is the reference knowledge adopted by most research works that will be reported in Section 3 and 4 The ODP, also known as Dmoz, is a multilingual open content directory of World Wide Web links. The ODP was founded in the United States as Gnuhoo by Rich Skrenta and Bob Truel in 1998 while they were both working as engineers for Sun Microsystems. It is now owned by Netscape, but it is maintained by a community of volunteer editors. The ODP uses a hierarchical ontology scheme for organizing web-sites listings. Listings on a similar topic are grouped into categories, arranged by subject (from broad to specific). Here below a small portion of categories listing obtained by the ODP archive (on date: 2010-05-11) is shown.

Adult
Adult/*Arts*
Adult/*Arts*/*Animation*
Adult/*Arts*/*Animation*/*Anime*
Adult/*Arts*/*Animation*/*Anime*/*Fan_Works*
Adult/*Arts*/*Animation*/*Anime*/*Fan_Works*/*Fan_Art*
Adult/*Arts*/*Animation*/*Anime*/*Fan_Works*/*Fan_Fiction*
Adult/*Arts*/*Animation*/*Anime*/*Games*
Adult/*Arts*/*Animation*/*Anime*/*Games*/*Companies*
Adult/*Arts*/*Animation*/*Anime*/*Games*/*Reviews*
Adult/*Arts*/*Animation*/*Anime*/*Genres*
...

To keep the ODP running smoothly, some set up policies for submitting sites are stated. Submissions that violate these policies are rejected. Furthermore, the ODP data are made available through a non standard RDF format. One strategy that allows to define a user profile is to automatically extract the relevant concepts from the preferred user's Web pages. *The ODP is the largest, most comprehensive human-edited directory of the Web*, thus many researchers have used it for their research activities.

3. The Use of Personalized Ontologies for Query Expansion

The aim of *query expansion* is to refine the user's query by adding new meaningful terms to the initial query in order to more effectively express the user's needs. The query refinement can be made either automatically or with the help of the user. In the literature, two main classes of methods for query expansion have been proposed [25]: *global* and *local* methods.

Global methods are independent from both the query and the results returned by the query evaluation. They include query expansion by using thesauri or the WordNet and query logs (particularly appropriate for Web search). With the support of a thesaurus each term t can be expanded with synonyms and words related to t. Thesaurus-based query expansion has the advantage of not requiring any user input, and generally it increases recall but decreases precision, especially when the query contains ambiguous terms [25]. Lin [26] is one of the pioneer researchers facing the problem of how the meaning of a phrase can change when one of the words is replaced by a similar word.

Local methods expand the query based on the information found in the result set produced by the initial query evaluation. They include relevance feedback (positive or negative relevant documents explicitly indicated by the user), pseudo-relevance feedback (or blind relevance feedback), and (global) indirect relevance feedback techniques. With relevance feedback the user indicates some of the results obtained after a first search as relevant or non-relevant. From the relevant results the significant concepts are extracted and added to the original query in order to refine the Web search. In pseudo-relevance feedback it is assumed that the top k ranked documents are relevant, and the relevance feedback is performed under this assumption. Instead, in the indirect relevance feedback *indirect* sources are used as the basis for relevance feedback. The area of research called *clickstream mining* [25] is a case study of indirect relevance feedback, and it is widely used on the Web [27,28]. If a user clicks on the links included on the Web pages, then it is assumed that those pages as relevant to him/her.

When using an ontology the concepts contained in it can be used for word sense disambiguation and subsequent query expansion. Ontologies offer a global technique of query expansion. The idea is to add terms of the ontology that are semantically related to query terms.

During the query expansion phase a crucial problem is to select the number of terms that have to be added to the query in order to better satisfy the user request [29]. What is the right number of candidate terms in the query expansion process? Two points of views are considered from researchers: a massive approach and a qualitative approach. In the first case, more than 20 terms can be added to the original query. Buckley [30] has indicated this methodology as the most suited one for routing the query vector towards the centroid of relevant documents. Whereas in the second case, 3-4 terms per query are usually inserted. Although a precise number is not indicated, the optimal number varies from query to query. The idea is that the number is less important than the quality of the selected terms [31]. In this scenario, ontologies can be used as a good support to the qualitative approach. The objective is to identify the right relationships in the user's ontology and to expand the queries with the right concepts.

Once a query term has been located in the ontology structure, two ways of navigation are possible in order to expand it: (1) broader navigation and (2) narrower navigation. By adopting the first approach a semantic generalization is considered, whereas with the second one a semantic specialization is analyzed. When using these techniques, the problem is to halt the expansion to the right level in the hierarchy [32,33] in order to avoid a massive approach.

In the literature, the query expansion process has been performed by adopting not only an ontological user profile, but also WordNet [34], EuroWordNet [35], MultiWordNet [36] or Cyc [37] as reference knowledge, or a specific domain ontology. WordNet (a lexical database for the English language) is an example of the most famous semantic network that has been widely used in the last few years in

numerous research works. In [33,38–43] some of the most important works that make use of WordNet in the query expansion process have been reported.

A specific ontology is domain dependent, and is defined by domain experts, for example we can have a Medical Ontology, a Tourism Ontology, an Architecture Ontology or an Animal Ontology. In [44–48] approaches that make use of a specific domain ontology for query expansion have been presented.

3.1. Approaches to Query Expansion Based on the Use of Personalized Ontologies

Query expansion aims to improve the user's Web search in order to retrieve more relevant documents. In this section, we review some approaches to query expansion where an ontology is semi-automatically defined on the basis of the observation of the user's behaviour during Web searches. In the literature, a limited number of such contributions can be found. In [32,49] query expansion strategies which consider both ontology and past user's queries are presented. In [32] the authors define a graph where each node is a query term, and with each term the related list of Web results is associated. This graph is updated each time a user's query is performed. Details on how to handle the exponential grown of this structure are not indicated in the paper; in fact, problems related to time and space complexity can arise. In [49], past user's searches are used to define a user's profile, and the ODP is used to define a general profile. These two profiles are then combined to establish the context of the user's query to personalize the current user's request. The authors assert that good results have been obtained with the defined strategy, but experimental results are performed only with seven users, a few hundred queries, and a limited number of relevant documents. Whereas in [32] past user's queries are involved only if there is not a direct correspondence between a user's query and the ontology, in [49] they are both jointly considered to better define the search context. In [50,51] local methods to expand the user's query including positive relevance feedback strategies are presented. Both these works define ontology-based query expansion methodologies applied in the e-learning context. Although [50] and [51] have the same goal, the first method is more complex from a computational point of view. Both methods define the user's ontology based on relevant terms extracted from e-learning material stored by the users. However, details on how this process is performed are missing. In the literature, the semi (or automatic) definition of a ontology from textual information is a challenging research topic [52]. As reported in Section 2 works that combine hybrid methodologies between query expansion and re-ranking phase are generally preferred. Thus, in this section we also discuss the approach defined in [53] where a hybrid combination of query expansion strategy for improving the re-ranking of Web results is presented. The terms to add to the query are directly selected from the ODP by the user, and the ODP is also used to extract the relevant concepts from the documents. In Table 1 a summary of the above approaches is reported.

In the following the papers previously introduced are more extensively explained.

In [32] the authors have defined a method for query expansion based on the use of ontologies. The use of an ontology reduces the possible (mis)interpretations of a query, especially when a user writes short queries. The authors have defined the concept of *ambiguity* as the gap between a user's query and his/her information need. In order to reduce this gap two types of ambiguity related to a query have been modelled: (i) *semantic ambiguity*, and (ii) *content-related ambiguity*.

Table 1. Synthesis of the discussed methods for query expansion.

Reference	Considered User's Actions	Personal Ontology
[32]	past user's queries	defined from textual information.
[49]	past user's queries	based on the ODP
[50,51]	preferred documents	defined from textual information.
[53]	selection of terms from the ODP	based on the ODP

The semantic ambiguity of a query can be evaluated when the query terms belong to the domain ontology. This means that a term x written in the query Q is associated with a concept c of the ontology, (*i.e.*, c is defined as $Type(x)$), and its relationships. The semantic ambiguity of a query Q is then defined as: $SemanticAmbiguity(Q) = \sum_{x \in Q} Variable_{Generality}(x) \cdot Variable_{Ambiguity}(x, Q)$, where x is a query term. The *variable generality* parameter, $Variable_{Generality}$, is the number of subconcepts (*i.e.*, children concepts) of c defined as $Variable_{Generality}(x) := |SubConcepts(Type(x))| + 1$, where the base value 1 allows to compute the value of ambiguity of the term x also when $|SubConcepts(Type(x))| = 0$. The *variable ambiguity* parameter, $Variable_{Ambiguity}(x, Q)$, is a number based on the relationships defined for the concepts identified for the query term x and based on the set of constraints related to x (for more details see [32]). A high value of *semantic ambiguity* indicates that the sub-concepts of $Type(x)$ could be good candidates to be added to the original query.

The content-related ambiguity is considered when there is not a direct correspondence between the query terms and the ontology structure. This means that the query terms are not represented in the ontology, and then it is not possible to evaluate the semantic ambiguity of a query. The content-related ambiguity is defined by calculating the so-called *Neighbourhood* of the query. The neighbourhood is a lattice defined by considering the relations between the queries' terms. This means that each query term is linked with the other terms written in the current user's query, and with the ones defined in past searches. After the user's query evaluation, each returned result is connected to each term in the query itself. This way a node of the lattice is a query term and a leaf is a result. When a user writes a new query, if the semantic ambiguity parameter cannot be computed, the lattice is analyzed in order to identify the set of results similar to the one obtained after the evaluation of the considered query. Two queries of a user are equivalent if their sets of results are equivalent. When the equivalent set of results has been identified in the lattice, it is then possible to locate the terms directly linked to these results. These terms are the candidates to be added to the original query.

In [49] a technique to improve the retrieval effectiveness of a search engine on the Web is presented. This solution makes use of two profiles: (1) a user's profile based on the user's search history, and (2) a general profile learned from a category hierarchy. These profiles are used to overcome the impersonal behaviour of a search engine given the context of a user's query.

A user's search history is represented by a tree model with the objective to capture the following information: queries, relevant documents, and related categories. In this tree model the root is a query, the nodes are the categories generated to classify the documents and the leaves are the relevant documents. The authors have assumed the use of search engines where the returned documents, after

the evaluation of a user's query, are classified into a set of categories (such as the Northern Light search engine). A document is considered relevant by a user if some of the following user behaviours are observed: a user selects it and reads it for a time slot, or a user saves/prints it. A user's profile is automatically learned from the user's search history, and it consists of a set of categories. For each category a weighted term vector is defined, where the term weight indicates the importance of the term for the category. A general profile is suitable for all users. The reason for using this additional information is that the knowledge acquired from a user is limited only to his/her previous requests, and thus it might not be sufficient to determine the appropriate context for a query by having new user's interests. The considered general knowledge for defining this profile is provided by the ODP. In particular, only the first three levels of the ODP's structure are taken into account to represent the set of all categories.

The personalized search is achieved by mapping a user's query to a set of categories (obtained after the analysis of the user and general profile) that define the query context. After this, a similarity value between the user's query and the relative set of categories is calculated; next the categories are ranked in descending order of similarity. The top three categories are visualized to the user who can select the best ones for his/her actual search. If the user's interests are not among these three top categories, then the system provides the next three ones, and so on. The query is initially submitted without specifying any category. Then, the query is again submitted by adding the selected categories to the initial query. These two lists of returned documents are merged by adopting a weighted voting-based merging algorithm; the combined list will be visualized to the user.

Lee *et al.*[50] have defined an algorithm aimed at disambiguating short user's queries. In the considered application the users have to locate e-learning material on the Internet. In this work the domain ontology has been semi-automatically defined by combining a domain vocabulary with the relevant concepts extracted from the preferred e-learning material. This information is stored in the repository after search sessions performed by users (e.g., students or teachers). A user is guided in locating the learning objects (e.g., books of interest) on the Internet by using titles, descriptions and other attributes.

The core of this approach is to identify the portion of knowledge related to the user's query in the considered domain ontology. The selected part of knowledge is a sub-tree of the ontology that the authors have called "User Intention Tree" (UIT). Thus, the UIT represents the user's interests that identifies the so-called *user intention* related to a given query. The original query is then expanded with the concepts of the UIT. In the ontology each concept c is described by a set of keywords and their synonyms, called *base concept*, BC, of c. Then for a given user's query Q the related UIT is identified by finding the right BCs base concepts. This phase is performed by assigning a score, (called the base concept score, BCS), to each concept as: $BCS(c_i) = \frac{|BC_i \cap Q|}{|BC_i|}\ \forall c_i \in C$, where C is the set of concepts of the ontology. The score of a selected concept c_i is defined as the number of keywords and synonyms that match Q (user query keywords) normalized by the total number of keywords and synonyms of concept c_i. For each pair of concepts directly linked in the ontology, with both having $BCS > 0$, a new score called weighted base concept score, $WBCS$, is calculated as: $WBCS(c_i, c_k) = BCS(c_i) \cdot BCS(c_k)$, where c_i is the considered concept and c_k is a concept directly connected with it having $BCS(c_k) > 0$. Then, the propagated base concept score, $PBCS$ for a concept c_i is defined as the sum over $WBCS$s. After

these phases, several sub-trees of the ontology (with weighted relations and concepts) are located. The last phase concerns the construction of the UIT. The several disjoint parts of the ontology (sub-trees) have to be aggregated to define the UIT. The aggregation process is defined by finding the common node containing a maximum conjunction of the base concepts from the sub-trees. As previously stated, the UIT concepts are added to the original query, but a great number of candidates can be obtained. In order to avoid a massive approach for the query expansion phase, two stages are defined for pruning the nodes. (1) In this step the degree of correlation between a concept c_i (having $BCS(c_i) > 0$), and its parent concept j is defined. For each c_j the total impact score (TIS) parameter is calculated as the sum of the impact scores, ISs, of all subconcepts (direct and indirect) having not null BCS value, *i.e.*, $TIS(c_j) = \sum IS(c_i, c_j)$. The impact score between the parent node and one of its child nodes is defined as: $IS(c_i, c_j) = \{PBCS(c_i) \cdot \frac{number\ of\ all\ descendants\ of\ c_i}{number\ of\ all\ descendants\ of\ c_j}\}$. If the concept is a leaf then the number of descendents is set to 1. The TIS value of each parent node j is compared with a threshold value γ to limit the upper bound of a minimal UIT. (2) If numerous nodes are obtained after the first stage, then a semantic distance between UIT concepts is defined to limit the expansion level. A threshold value is assumed to discard the concepts with a low distance value. This approach could be onerous with respect to the computational time if the ontology is made up of a large number of concepts.

In e-learning systems users have to be able to easily discover the relevant information which supports learning activity. To enhance the performance of the searching task, query expansion may be used. In [51] a query expansion methodology based on a semantic user model that represents individual user's interests has been presented. A user model is exploited to represent the interests and background knowledge of individual learners [54]. Formally, a user model $U_{model} = (UP, IC)$ consists of a set UP of user's identity information (such as user's name, age, education level, learning orientation, ...), and a set IC describing user's interests. The representation and the construction of IC is the main part of U_{model}. In the considered e-learning system each resource is annotated with metadata to describe its content, and a domain ontology is used to represent the concepts, instances and relationships from those annotations. The IC set is defined by considering the k resources selected by a user. From the obtained k resources the metadata are extracted, and put together with the keywords defined by a user to formulate his/her query. Formally we have $W = UK \cup MD$, where UK is the set of the keywords from the user's query, and MD is the set of all the metadata annotations from the k resources. For each $w_i \in W$ the corresponding set of synonyms, SYN, is obtained by analyzing a specific domain lexicon. From SYN the related concepts, instances and relationships of the domain ontology are extracted to construct the set of user's interests IC. Once the IC set is defined, the process of query expansion is obtained in several steps: (1) from each query term its synonyms are identified in a specific domain lexicon, (2) the query terms and their synonyms are used to mine concepts, instances and relationships from the domain ontology, (3) the extracted ontology obtained from point (2) is compared with the specific user model IC to identify the interests related to the current query, (4) a similarity measure is used to arrange the elements extracted from IC in descending order of corresponding interest relevance degree, and (5) the top k concepts (or instances) are added to the query. Evaluations have shown good results by comparing documents obtained with and without the query expansion strategy. Furthermore, the behaviour of two users with the same queries but having a different user model IC has been compared. Evaluations

have produced a different set of returned search results, thus providing a personalized service in the e-learning system.

If two people are searching for *raptors* they may be looking at different contexts (prehistoric animals or basketball teams), but they will obtain the same set of results. To address this problem, in [53] a search engine called *KeyConcept* has been developed. This search engine takes into account, in addition to the query terms, the concepts related to the query and selected from the ODP. In particular, the user can directly select the ODP concepts most related to the keywords written in the query with the help of an ad-hoc interface. Furthermore, the documents returned after the evaluation of a query are indexed by their keywords and by their concepts again obtained from the ODP. Thus, documents are re-ranked according to a match between the user-supplied keywords/concepts and those associated with the documents. In detail, for each document two scores are calculated representing the number of keywords/concepts matching those specified by the user, called $Score_k$ and $Score_c$, respectively. A user, during the definition of the query, can specify an α parameter, between 0 and 1, in order to establish if during the computation of the document score, the $Score_c$ parameter can assume more importance than the $Score_k$ parameter. The document score, $Score_d$, is computed as: $Score_d = (\alpha \cdot Score_c) + ((1-\alpha) \cdot Score_k)$. Once a score has been calculated for each document, then they are sorted in decreasing order and presented to the user. To avoid having a great number of documents, the hierarchical structure of the ODP is explored in order to prune the result set. In particular, the document that does not share the nodes belonging to the same direct path of the ODP structure with respect to the user-supplied concepts is removed. This way the result set is made up of documents related to the user's request.

3.2. Future Directions for Query Expansion Based on Personalized Ontologies

In the literature, a wide number of query expansion strategies has been defined. In Section 3 some approaches based on an ontological user profile to query expansion have been presented.

In the query expansion strategy, a possible future investigation is to consider multiple ontologies to create a tailored user's profile. Several works have analyzed the joint usage of a domain ontology with a global ontology. For example, in [55] the authors have taken into account the problem of expanding queries with terms satisfying spatial constraints. To this aim a tourism ontology and a global ontology with geographic information are used. In [56] the authors present *Miology*, a Web-based application that helps users to automatically select the right ontologies (from an ontology repository) related to the topics of given Web results. Future investigations can be addressed to the use of these topic-ontologies for adding more focused terms to the user's query.

Another future direction could be to allow users to interact with the knowledge model (the ontology in this context) to increase the effectiveness of query expansion techniques [57]. In [58] a study about the effectiveness of interactive query expansion within the context of a relevance feedback system has been performed. In this study, it emerges that about one-third of the terms directly chosen by the users from a list of candidate terms is potentially useful to improve the phase of query expansion. For example, in [59] the Concept based Information Retrieval Interface (CIRI) ontology has been developed to help the user in formulating his/her queries. The experiment was a pilot study of the Finnish Web ontologies project into the *Food industry* domain. By using CIRI, a user can navigate the ontology by choosing the concepts up to the desired number of expansion levels. A promising research topic could be to make

use of personal ontologies to better exploit an interactive search task. This *navigational query expansion* allows the user to have an active role during her/his searches. In [60] the authors define the so-called ontology query model to produce an Information Retrieval query from a set of concepts selected by a user browsing the considered ontology. Experiments are performed on a biomedical ontology merging several well-known biomedical sources, such as the Gene Ontology, MeSH and Swissprot databases.

4. The Use of Personalized Ontologies for Re-ranking Search Results

After a Web search, the produced results are usually ranked according to a numerical estimate of the document relevance to the query. This order does not always conform to the real user preferences. The phase of re-ranking allows to re-arrange the retrieved information items in order to help the user in locating those relevant to his/her needs. In this section we analyze the approaches that make use of an ontological user profile to re-rank the results produced by a query evaluation. An ontological user profile is generally defined by extracting information from the relevant Web pages obtained in previous user searches. We analyze works where a Web page is considered relevant based on an analysis of actions performed by a user on it. In the considered approaches (*i.e.*, [10,21,61–63]) the usual actions taken into account for selecting the preferred Web pages are: the frequency of the visits to a page, the time spent on a page, bookmarking or saving a page, etc. The relevant concepts are then extracted from the preferred Web pages and linked together with the support of external knowledge, such as the ODP. In [9] an approach analogous to the previous ones is presented; the monitoring of the user's activities has been performed by implementing a wrapper around the Google search engine instead of considering proxy servers or desktop bots. In [13,64] a user profile is also defined with the direct interaction of a user with the system. In both cases a user can specify his/her categories of interest: in [64] this is done by navigating the concepts and selecting the right ones from the ODP, whereas in [13] the categories are explicitly declared. Table 2 reports a summary of the above approaches.

Table 2. Synthesis of the discussed methods for re-ranking of search results.

Reference	Considered User's Actions	Personal Ontology
[10,21,61–63]	frequency of the visits to a page, time spent on a page, saving a page, *etc.*	based on the ODP
[9]	past user's queries	based on the ODP
[64]	selection of terms from the ODP	based on the ODP
[13]	past user's queries, the preferred documents and interests explicitly declared.	based on the ODP

4.1. Approaches Defined to Re-rank the Search Results Based on the Use of Personalized Ontologies

In [10] the authors have defined a Web system which aims to improve the quality of a user's search on the Web. The objective is obtained by personalizing a Web search according to the user's interests. The authors state that a personalized search can be of two types: *context oriented* and *individual oriented*. In

the first case, the context of a search can be derived from the query terms written by a user; whereas in the second case, the context of a search is obtained by analyzing the user's behaviour during previous Web searches, e.g. actions such as saving, printing or copying full or part of the viewed pages. In the Web system developed by the authors both these types of context are taken into account to improve the quality of a search. The results obtained by the existing search engines, after the evaluation of a user's query, are collected, analyzed, re-ranked based on the user's interests, and then recommended to the users through the given user interface. In particular, in this work two types of ontologies can be used to personalize a Web search: a page ontology and a personalized ontology. Page ontology is an incremental ontology, it is constructed based on the set of pages visited by the user during his/her search session; whereas a personalized ontology is defined by considering several factors of personalization, such as page view time, actions performed on a page, terms written in previous queries, and the relevant concepts extracted from the examined Web pages.

The page ontology is generated to exploit the links between the various pages visited by the user after a search session. From each visited page the relevant concepts are extracted, and they are used to represent such pages. Relations between concepts are defined with the adoption of both the ODP and a dictionary. The re-ranking process privileges the pages having the highest number of concepts related to the user's query, and where the user spent the most time. These pages will have a higher ranking position than the other ones.

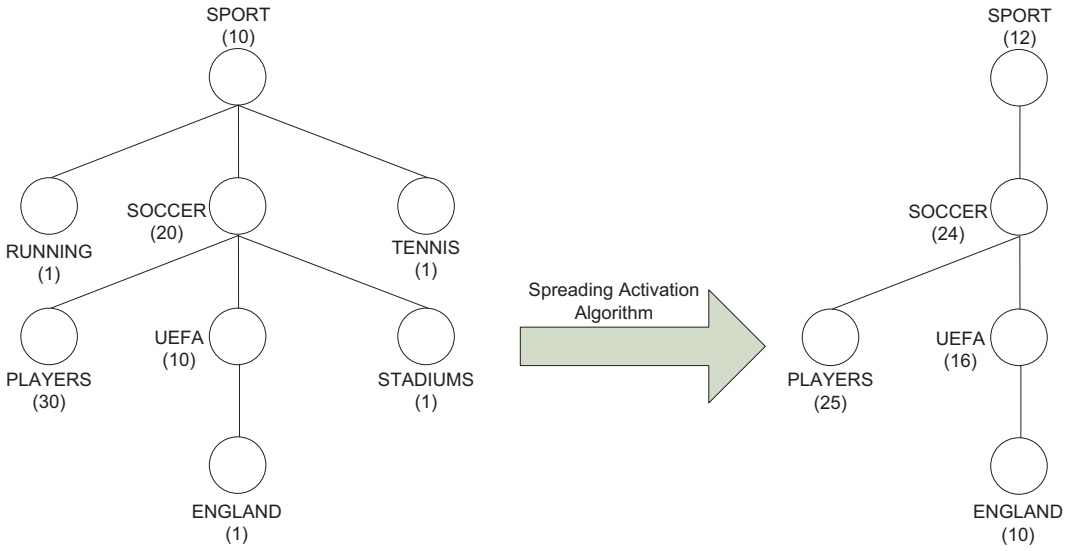
The personalized ontology is defined by performing a mapping between the relevant concepts extracted from the visited pages and the concepts of interest extracted from the search queries. The extracted concepts are linked in a conceptual way with the support of both the ODP and a dictionary. In particular, $is-a$ and $part-of$ are the relationships identified between the concepts in this step. Each node in the personalized ontology is defined by three components: a concept, the Web page(s) containing the concept, and a weight. The weight is calculated based on the actions performed by the user on this particular Web page. Each action determines the page relevance with respect to the user's context of search. For example, if $Action = Save$ then the $weight = 1$ or if $Action = Copy$ then weight will have a value from 0.25 to 0.75. The Protégé tool can define the personalized ontology in a specific ontological language. The pages having concepts with a high weight will obtain a high position ranking.

The goal of a personalized Web search is to tailor search results to the user's interests and preferences. In [21,61,62] a strategy to define an ontological user profile to personalize search results for a given query has been presented. The personal ontology is built by considering the relevant concepts extracted from the preferred Web pages after a Web search. Furthermore, a score is assigned to each concept. Initially all scores are set to 1. The consideration of this score plays a key role in the re-ranking process. As a first step, the Web pages are collected for which the user has shown interest. The factors taken into account in order to select the preferred pages from all the ones viewed are: the frequency of the visits to a page, the amount of time spent on the page, and other user actions such as bookmarking a page, etc. After a user's query, the preferred documents are then analyzed, and the more relevant concepts are extracted from them. In particular, each preferred document is represented by a vector $\bar{d}_i$ of its relevant concepts. In this work, the ODP is also used in order to link the concepts extracted from the visited pages in a hierarchical way. Each node in the ontology is a pair $< C_j, IS(C_j) >$, where C_j is the concept in the reference ontology, and $IS(C_j)$ is the interest score for that concept. Furthermore, for each concept C_j a vector $\bar{C}_j$

is defined, which contains all the sub-concepts (direct children nodes) of C_j. When the relevant concepts of each document vector $\bar{d}_i$ are located in the reference ontology, their scores are updated according to two algorithms presented in [21,61,62]. From the first algorithm the *activation value* for each concept is calculated in order to define the importance of the concept for the current user's query. The activation value of a concept is defined as: $activation(C_j) = sim(\bar{d}_i, \bar{C}_j) \cdot IS(C_j)$, where $sim(\bar{d}_i, \bar{C}_j)$ is the cosine similarity measure calculated for two vectors $\bar{d}_i$ and $\bar{C}_j$, while $IS(C_j)$ is the existent interest score for the specific concept stored in the ontology. From the second algorithm the score of the concept C_j is updated as defined in [21,61,62]. This process is performed after each user's query, so that when several searches are performed by the user, a sub-structure from the ODP is obtained by considering the relevant concepts. This sub-structure is related to the user's interests, and it is used in order to re-rank the search results returned after the evaluation of a given query. In detail, for each result the relevant concepts are extracted and weighted according to the formula described in [61]. Then the similarity of the document with each concept of the sub-structure is computed to identify the best matching concept. Once the best matching concept, C_j, is located, a rank score is assigned to the document by multiplying the $IS(C_j)$, the similarity of the document to the query and the similarity of the specific concept to the query. Once all documents have been analyzed, then the search results are re-ranked in decreasing order and presented to the user.

Figure 2 shows an example of the updating of scores assigned to the concepts after the application of the algorithm. Let us suppose that a Web result of interest is related to the concepts $uefa$ and $england$. The algorithm computes a new higher score for the concepts $sport, soccer, uefa$ and $england$, *i.e.*, for the concepts where the user has shown interest and for those belonging to the same direct path of the hierarchy; a new lower score is associated to the other sub-concepts, *i.e.*, $players$. As stated before, the minimum weight assigned to each concept is 1, so that in this example (after the application of the algorithm) the concepts having a score equal to 1 are not reported.

Figure 2. Schema of a small part of an ontological user profile where the scores to the concepts are updated after a user's search (see [21]).



In [63] during a search session defined as a sequence of related queries, a user profile is built which refers to the user's interests in the search session. The user profile is initialized after a first user's query and updated after the others user's queries. Each time that a user submits a query to the system, the search results are re-ranked according to the user profile defined up to the previous queries. In detail, a search session S is defined at time $t+n$ as a sequence of related search activities performed by queries $\{q^0, \ldots, q^{t-1}, q^t, q^{t+1}, \ldots, q^{t+n}\}$ submitted respectively at time $\{0, \ldots, t-1, t, t+1, \ldots, t+n\}$. In order to consider queries related to the same search context, the authors have proposed a session boundary recognition method by using the *Kendall* rank correlation measure that quantifies the correlation between the current user query (e.g., defined at time t) and the user profile (e.g., defined until time $t-1$). A threshold value σ is used in order to define if a query defined at time t is related to the same context of the previous queries; *i.e.*, this is verified only if the value obtained after the calculus of the Kendall measure is greater than σ.

If the query q^t submitted at time t is related to the some previous search activities, then the documents of interest returned in response to the user's query are collected. A document is relevant, for a user, if some actions on that document are performed (such as, saving, printing, coping,...). Once all the documents of interest have been identified, D^t , then the authors have defined the query context, K^t, as the centroid of the documents in D^t; namely the weight of each term, $term_i$, in K^t is calculated as $K^t(term_i) = \frac{1}{|D^t|} \cdot \sum_{d_k \in D^t} w_{term_i d_k}$, where $w_{term_i d_k}$ is the weight of the $term_i$ in document d_k. At this point a vector $\bar{K}^t$ is obtained having the most relevant (weighted-)terms from the documents returned after the evaluation of the query.

A user profile is represented as a weighted graph where the nodes are the categories (or concepts) defined in the ODP. This means that concepts are linked by using the taxonomic relationship IS-A. To re-rank the documents according to the user profile, it means that only the concepts that represent the Web pages (for more details see [65]) visited by the user during the previous queries at time $t-1$ are taken into account. Then, each selected concept from the ODP (*i.e.*, c_j) is represented as a single term vector $\bar{c_j}$, and its similarity score with $\bar{K}^t$ is computed as $score(c_j) = cos(\bar{c_j}, \bar{K}^t.)$. When the scores of the concepts are updated, the D^t set is analyzed and compared with the user's profile in order to re-rank the documents according to the user interests.

For each document, d_k, obtained after the evaluation of a query at time t, a weight is computed as: $W(d_k) = \alpha \cdot S_i(q^0, d_k) + (1-\alpha) \cdot S_c(d_k, G^t)$ for $0 < \alpha < 1$, where S_i is the similarity value between the first user's query and the document d_k, while S_c is the similarity value between the document d_k and the user profile at time t. Both the similarity values are calculated by adopting the cosine similarity measure. Once the results are analyzed, then this weight is used in order to re-rank them in decreasing order.

In several approaches user profiles are created by considering proxy servers (e.g., to capture browsing histories) or desktop bots (e.g., to capture activities on a personal computer). In [9] an approach is proposed which implements a wrapper around the Google search engine. In this work, a user profile is analyzed to identify the user's interests related to a query in order to improve the quality of a user's search. User profiles are created by classifying the information collected during previous search sessions into concepts with the support of a reference hierarchy (*i.e.*, the OPD). In detail, the collected information is related to queries and snippets (titles and summaries) of the results. *These profiles are then used to*

re-rank the search results and the rank-order of the user-examined results before and after re-ranking have been compared [9].

User profiles are defined by considering the reference concepts from the ODP. Each concept (or category) of the ODP is weighted, where higher values imply major interest for this category. Each weight is assigned by classifying the textual content collected from the user in the right category. This means that the relevant concepts extracted from the information stored by the user are classified in the ODP hierarchy. This procedure produces a list of concepts with an associated weight that can be updated (or defined) over time after the submission of user's queries, or after the analysis of preferred snippets by the user. The formulae of how to calculate the weight for each concept are given in [66].

The re-ranking phase is performed in two steps. Step (1): for each search result obtained after the evaluation of a user's query, the relevant concepts are extracted and located in the ODP. Furthermore, for each category of the ODP a weight is calculated according to the formula defined in [66]. This process allows to define the so-called *Web profile* produced by the classification of each Web result. Then, the conceptual match between the Web profile of the $j-th$ result (*i.e.*, res_j), and the user profile (*i.e.*, $user_i$) is defined as: $conceptual\_match(user_i, res_j) = \sum_{k=1}^{N} cwt_{ik} \cdot cwt_{jk}$, where cwt_{ik} is the weight of concept k in the user profile i, cwt_{jk} is the weight of concept k in the Web profile j, and N is the number of concepts. This new value is called conceptual rank. Step (2): the final rank of the $j-th$ result is calculated by combining the conceptual rank with Google's original rank as: $FinalRank(res_j) = \alpha \cdot ConceptualRank_j + (1-\alpha) \cdot GoogleRank_j$, where α has value between 0 and 1. If $\alpha = 0$, then the conceptual rank is not considered, and the $FinalRank$ is equal to the original rank provided by Google. If $\alpha = 1$, then search engine ranking is ignored, and only the conceptual rank is considered. The authors have performed evaluations based on two user profiles (for a same user) by considering two types of information: past user's queries and snippets obtained in response to a user's query, respectively. Evaluations on these user profiles are effective in the same way.

In [64] a user profile based on the ODP is used in order to re-rank the Web search results. In this work, a user profile is directly defined by the user. In fact, a user has to select the topics that best fit his/her interests from the ODP. The ODP categories (topics of interest) are organized in a hierarchical way. When a user has chosen the categories of interest, the path of each category is stored in the user profile, *i.e.* all categories until the root. For example, a part of a user profile could be:

$$Music/Styles/Jazz,\ Music/Instruments/Guitar,\ Music/Artist/Lewis.$$

Once the user profile has been defined, then the output given after the evaluation of a query is re-ranked by computing a distance between the user profile and each output result. In detail, after the evaluation of a user's query, the obtained results are stored and then compared with the user profile in order to establish which result is most related to the user's interests. For each result the relevant concepts are related to the ODP categories. As a consequence, a result is classified according to the ODP ontology as the user profile. This way the distance between a user profile and a result, $Dist_j$, is performed by analyzing their distance in the ODP structure. The ODP can be considered as a tree, and its categories/topics can be nodes. From this point of view, the distance is related to two sets of nodes from the topic tree: (1) those representing the user profile (set A), and (2) those associated to the result (set B). The distance between these sets is established as the minimum distance between all pairs of nodes given by the Cartesian

product $A \times B$. Thus, the pair of nodes having the minimum distance is considered in order to calculate the $Dist_j$ score to associate to each result. This score is computed by combining the 'Complex distance' (see [64]) between a and b, and the Google PageRank score of b; where node a identifies the user profile, and node b is the result. Once all the results are processed, then it is possible to re-rank them in decreasing order according to their distance value.

Several experimental evaluations have been performed by adopting three types of queries (clear query, relatively ambiguous query, ambiguous query) in different types of Web searches, such as a personalized ODP search and a personalized Google search by using the Google Directory as reference taxonomy (where Google Directory is a sub-structure of the ODP) instead of the ODP. Thus, by considering the metric previously presented, a user profile is compared with the ODP taxonomy and the Google directory, respectively. In detail, 17 users have defined their profile by selecting the categories of interest from the ODP, and then they have defined their queries. The top 5 results obtained after the evaluation of a query have been re-ranked by applying the above approach on two reference taxonomies, *i.e.*, the ODP and the Google Directory. Thus, for a same query two different rankings of the same result set have been obtained. By using the ODP, good results are obtained for the following types of user's queries: relatively ambiguous query and ambiguous query.

In [13] a personalized information retrieval model based on ontological knowledge has been proposed. The ontology, in this case, is built by considering two types of user's interests: persistent and live. Persistent interests are related to the preferences explicitly declared by the user, whereas the live interests are obtained by monitoring the user's actions during a search session. These two methods allow to jointly define the user's profile. To define a persistent interest, a user selects a-priori a set of different topics with a certain degree. The live interest is generated through a constant monitoring of the user's interaction with the system during a search session. As stated by the authors, this last user interest contains less uncertainty because it is based on his/her actions, as long as the monitoring period is sufficient and representative of the user's preferences [13]. The live user profile is defined according to four actions: *keyword-based queries, view document queries, relevance feedback queries, and browsing queries*. *Keyword-based queries* define a fuzzy set of concepts on the keywords extracted from the user's queries, *view document queries* define a fuzzy set of concepts extracted from the visited documents by the user, *relevance feedback queries* denote two sets of documents: one is the set of documents marked by the user as relevant, and the other is the set of documents marked by the user as non-relevant. At the end, *browsing queries* define a fuzzy set of topics requested for browsing by the user.

Hence the ontology is built by jointly considering all these sets from the live interests, plus the preferences assigned directly from the user (*i.e.*, persistent interest). The ontology knowledge is then defined with the adoption of a formal methodology founded on fuzzy relational algebra [67,68]. This formal methodology allows to extract the relevant concepts from the user's interests (persistent and live) and to link them. This way, two concepts are linked, but the semantics of this relation is not known. The semantics is given by the domain expert from the following set of taxonomic relationships $\{specialization, part, example, instrument, location, patient, property\}$, with the corresponding set of inverse relations.

This way each user's request is compared to the ontological user profile. The ontology is analyzed in order to discover the part of knowledge related to the user's query. As a consequence a different ranking of the results based on the user's interests is computed.

Up until now in the works outlined in this paper the relevant concepts from the documents are extracted and semantically linked with the support of a reference ontology (*i.e.*, the ODP). Instead, in this case the ontology is built without any external support and in a semi-automatic way: a formal methodology is applied in order to link the concepts, and the domain expert defines the semantics of the links.

To annotate documents with concepts and to define the corresponding ontology without an external support is a very difficult process. In the literature, several criticisms have emerged towards the use of formal annotation and ontologies for capturing the semantics of documents [69–71]. An ontology is both domain and language dependent, and of course, it is very hard work to define an ontology representing the knowledge carried by a document or a Web page.

4.2. Future Directions for Re-ranking Based on Personalized Ontologies

In the presented approaches to re-rank search results, the trend line is to adopt an external reference knowledge as a semantic support to define an ontological user profile based on the concepts extracted from the relevant Web pages. The ODP is the knowledge used for this aim. Although the ODP is constantly updated and maintained by a community of volunteer editors, the subsumption relation (*i.e.*, the taxonomic IS-A hierarchy) between categories is the only semantic information which is provided. A further step could be to use a real general purpose ontology as reference knowledge where a richer semantic expressiveness is given. YAGO is a large ontology, automatically obtained from Wikipedia and WordNet with high coverage and precision [72,73]. YAGO has been defined in the last two years and could be a good candidate as a support to personalized searches. Currently YAGO contains more than 1.7 million entities and 15 million facts, grouped by 99 relations. In YAGO a fact is an instantiated relation between entities. Among the several relations, the IS-A is included as well as additional semantic relations between entities. In the literature, some works that make use of YAGO are proposed. For example, in [74], YAGO is used in order to classify the search results into appropriate categories. The goal of this approach is not to re-rank the search results according to the user profile, but given a user's query, the goal is to provide all the categories for the corresponding search results. This way a user is helped in locating the Web results by navigating the obtained categories' list.

A promising research direction can address the use of NAGA [75,76]. NAGA is a new semantic search engine, and it can operate on knowledge bases organized as graphs. NAGA searches the sub-graphs that match the user's query in the knowledge base, and these sub-graphs are the results obtained by the user as response. NAGA's goal is to rank the retrieved answers in such way that the most important answers are ranked before. YAGO is the knowledge used to perform preliminary experiments in order to compare the quality of the search results provided by NAGA with the ones obtained by a search engine, such as Google, and the question answering systems, such as Yahoo! Answers [77] and START [78]. In [79] the authors have proposed an approach to personalize the user's search in NAGA. The objective is to establish a user profile as several sub-graphs into NAGA and to define a personalized scoring model into the NAGA's ranking function that makes use of the user's interests. A user's profile is directly elicited

by the user himself/herself by querying YAGO's knowledge with NAGA for discovering his/her interests from the obtained results (*i.e.*, the related sub-graphs). To facilitate this phase, an ad-hoc interface has been developed, but this limits the interoperability of the approach where a user can interact with the system only by formulating queries with an appropriate interface and a specific query language. As previously stated, YAGO has a richer semantic than the ODP, but YAGO is a static general purpose ontology, whereas ODP is always updated from the community of volunteers. Future investigations can be made in two directions: (1) to define a user profile based on YAGO, and (2) to study new strategies for updating YAGO's knowledge.

5. Conclusions and Future Works

Web search engines provide a huge number of answers in response to a user query with the consequence that a user cannot always find the results relevant to his/her information needs. In order to overcome this unsatisfying situation, a possible solution is to analyze the behaviour of a user during a search session, namely (1) to study the interactions of a user with search engines, and/or (2) to study the actions that he/she performs visiting Web pages. With the first case, the terms of a user's queries are stored into query log files; whereas in the second case, information such as the identifier of a user, URLs clicked for each query or actions such as save, print, copy, etc. related to a part of a Web page are stored in Web log files. All these files are analyzed in order to understand what the user's interests are, and thus to define ad-hoc the profile. This profile is used to improve the quality of a user's search. In this paper, we have investigated the use of ontologies, aimed at defining user profiles (called *ontological user profiles*). Ontologies allow to give a semantic organization to the information recorded in query and Web log files. In this short review, some works aimed at the use of a personal ontology have been discussed in order to face two typical Information Retrieval problems: query reformulation and results re-ranking. Most works use ODP as the reference ontology to define a user profile. The ODP is used in three different ways. In the first case, ODP has been used as a semantic support to find a relation between concepts (e.g., extracted from the preferred documents or the visited Web pages), in the second case, some parts of the ODP structure have been identified to be relevant for the user, and in the last strategy, the user has directly selected the categories (concepts) of interest from the ODP.

The benefits of adopting strategies to obtain personalized searches on the Web are well known. In this paper, our interest has been directed towards two main approaches, namely query expansion and result re-ranking, to personalize a user's search with the support of personalized ontologies. With the coming of the Web 2.0, the way to share and define data has changed. New services are provided to handle folksonomies, blogs, wikis, and so on. In the literature, several works have addressed the problem of exploring folksonomies for personalized search. The term *folksonomy* is a combination of *folk* and *taxonomy* to describe the social classification phenomenon [80]. Folksonomy arises from the textual annotation of Web resources and is described as a shared collection of tags. A tag is a *free-text keyword*, and tagging is the process aimed at assigning tags to resources. As free-text keywords, tags do not have exact meaning and suffer from linguistic ambiguities. In current folksonomies there is no agreement on the representation on such a tagging. This means that each system uses a different format to publish its tagging data with the consequence of having a difficult interoperability between them. In [81] an overview and a comparison of several tag ontologies defined with the objective to have a

unique semantic representation has been presented. In this field of research, the main objective is to propose new approaches for improving a user's search. In [80,82,83] strategies for re-ranking search results are presented, based on the definition of a user profile, obtained after the analysis of bookmarked documents. In particular, a user profile is defined with the tags associated to each bookmarked document with the basic assumption that a user bookmarks only relevant interests to him/her. Metadata about Web results are collected by the community of users submitting bookmarks to the bookmark systems defining the so-called *document profile*. The re-ranking phase is then obtained calculating the similarity between a user profile and a document profile.

Recommendation system is an other interesting area of research where studies for the definition of a user profile from folksonomy are investigated. For example, in [84] it has been observed that users generate several tags spanned across many different domains. This means that users usually have a wide range of interests. To understand what the user's interests are, the distribution on how tags are assigned to bookmark documents has been studied.

The use of ontological user profiles obtained by folksonomies can improve the personalized search of the Web 2.0. Ontologies can be used also in another direction with respect to the one to give a common representation of tags. In fact, the process of tagging is nothing more than annotating Web resources with an unstructured list of tags, and all the limits concerning the semantic interpretation of tags are well known.

References

1. Shahabi, C.; Chen, Y.S. Web Information Personalization: Challenges and Approaches. *Lect. Note. Comput. Sci.* **2003**, *2822*, 5–15.
2. Micarelli, A.; Gasparetti, F.; Sciarrone, F.; Gauch, S. Personalized Search on the World Wide Web. *Lect. Note. Comput. Sci.* **2007**, *4321*, 195–230.
3. Lawrence, S. Context in Web Search. *IEEE Data Eng. Bull.* **2000**, *23*, 25–32.
4. Jing, B.; Jian-Yun, N.; Guihong, C.; Hugues, B. Using query contexts in information retrieval. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 2007; ACM: New York, NY, USA, 2007; pp. 15–22.
5. Wilson, T.D. Human information behavior. *Inf. Sci.* **2000**, *3*, 49–56.
6. Bruce, H.; Jones, W.; Dumais, S. Information behaviour that keeps found things found. *Inf. Res.* **2004**, *10*, 1–22.
7. Hepworth, M. Knowledge of information behaviour and its relevance to the design of people-centred information product and services. *J. Doc.* **2007**, *63*, 33–56.
8. Open Directory Project. Available online: http://dmoz.org/ (accessed on 2 March 2010).
9. Speretta, M.; Gauch, S. Personalized search based on user search histories. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Compigne, France, 19–22 September 2005; IEEE Computer Society: Washington, DC, USA, 2005; pp. 622–628.
10. Sendhilkumar, S.; Geetha, T.V. Personalized ontology for web search personalization. In Proceedings of the 1st Bangalore annual Compute conference, Bangalore, India, 18–20 January 2008; ACM: New York, NY, USA, 2008; pp. 1–7.

11. Silverstein, C.; Marais, H.; Henzinger, M.; Moricz, M. Analysis of a very large web search engine query log. *SIGIR Forum* **1999**, *33*, 6–12.
12. Vallet, D.; Fernández, M.; Castells, P.; Mylonas, P.; Avrithis, Y. Personalized Information Retrieval in Context. In Proceedings of the 3rd International Workshop on Modeling and Retrieval of Context, Boston, USA, 16–17 July 2006; AAAI Press: Boston, MA, USA, 2006; pp. 1–5.
13. Mylonas, P.; Vallet, D.; Castells, P.; Fernandez, M.; Avrithis, Y. Personalized information retrieval based on context and ontological knowledge. *Knowl. Eng. Rev.* **2008**, *23*, 73–100.
14. Suchman, L.A. *Plans and Situated Actions: The Problem of Human-Machine Communication*; Cambridge University Press: New York, NY, USA, 1987.
15. Lammari, N.; Mtais, E. Building and maintaining ontologies: A set of algorithms. *Data Knowl. Eng.* **2004**, *48*, 155–176.
16. Gruber, T. A Translation Approach to Portable Ontology Specifications. *Knowl. Acquis.* **1993**, *5*, 199–220.
17. Guarino, N.; Giaretta, P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases, Enschede, Netherlands, April 1995; Mars, N., Ed.; IOS Press: Amsterdam, Netherlands, 1995; pp. 25–32.
18. Tamma, V. An Ontology Model Supporting Multiple Ontologies for Knowledge Sharing. PhD Thesis, University of Liverpool, Liverpool, UK, October 2001.
19. Soo, V.W.; Lin, C.Y. Ontology-based information retrieval in a multi-agent system for digital library. In Proceedings of the 6th Conference on Artificial Intelligence and Applications, Kaohsiung, Taiwan, 9 November 2001; CPS: Taipei, Taiwan, 2001; pp. 241–246.
20. Bordogna, G.; Pasi, G. A flexible multi criteria information filtering model. *Soft Comput.* **2010**, *14*, 799–809; DOI: 10.1007/s00500-009-0476-3.
21. Sieg, A.; Mobasher, B.; Burke, R. Ontological User Profiles for Representing Context in Web Search. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology—Workshops, Silicon Valley, USA, 2–5 November 2007; IEEE Computer Society: Washington, DC, USA, 2007; pp. 91–94.
22. Guerrero, C.; Juiz, C.; Puigjaner, R. Web performance and behavior ontology. In Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, Barcelona, Spain, 4–7 March 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 219–225.
23. Gauch, S.; Chaffee, J.; Pretschner, A. Ontology-based personalized search and browsing. *Web Intell. Agent Syst.* **2003**, *1*, 1–3.
24. Natsev, A.P.; Haubold, A.; Tešić, J.; Xie, L.; Yan, R. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 24–29 September 2007; ACM: New York, NY, USA, 2007; pp. 991–1000.
25. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
26. Lin, D. Automatic Identification of Non-compositional Phrases. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics,

Baltimore, Maryland, 13–15 October 1999; Association for Computational Linguistics: Morristown, NJ, USA, 1999; pp. 317–324.

27. Department, P.C.; Chatterjee, P.; Hoffman, D.L.; Novak, T.P. Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Market. Sci.* **2000**, *22*, 520–541.
28. Montgomery, A.L.; Li, S.; Srinivasan, K.; Liechty, J.C. Modeling Online Browsing and Path Analysis Using Clickstream Data. *Market. Sci.* **2004**, *23*, 579–595.
29. Ogilvie, P.; Voorhees, E.; Callan, J. On the number of terms used in automatic query expansion. *Inf. Retr.* **2009**, *12*, 666–679.
30. Buckley, C. Automatic Query Expansion Using SMART : TREC 3. In Proceedings of The Third Text REtrieval Conference (TREC-3), Gaithersburg, Maryland, 2–4 November 1995; pp. 69–80.
31. Sihvinen, A.; Vakkari, P. Subject knowledge improves interactive query expansion assisted by a thesaurus. *J. Doc.* **2004**, *21*, 475–487.
32. Stojanovic, N.; Studer, R.; Stojanovic, L. An Approach for Step-By-Step Query Refinement in the Ontology-Based Information Retrieval. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, 20–24 September 2004; IEEE Computer Society: Washington, DC, USA, 2004; pp. 36–43.
33. Voorhees, E.M. Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 3–6 July 1994; Springer-Verlag: New York, Inc.: New York, NY, USA, 1994; pp. 61–69.
34. WordNet. Available online: http://wordnet.princeton.edu/ (accessed on 13 March 2010).
35. EuroWordNet. Available online: http://www.illc.uva.nl/EuroWordNet/ (accessed on 13 March 2010).
36. MultiWordNet. Available online: http://multiwordnet.itc.it/english/home.php/ (accessed on 13 March 2010).
37. Cyc. Available online: http://www.cyc.com/ (accessed on 13 March 2010).
38. Navigli, R.; Velardi, P. An analysis of ontology-based query expansion strategies. In Proceedings of the Workshop on Adaptive Text Extraction and Mining, Cavtat Dubrovnik, Croatia, 22–26 September 2003; Available online: http://www.dcs.shef.ac.uk/ fabio/ATEM03/ ATEM03-Proceedings.pdf/ (accessed on 18 October 2010).
39. Song, M.; Song, I.Y.; Hu, X.; Allen, R.B. Integration of association rules and ontologies for semantic query expansion. *Data Knowl. Eng.* **2007**, *63*, 63–75.
40. Liu, S.; Liu, F.; Yu, C.; Meng, W. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004; ACM: New York, NY, USA, 2004; pp. 266–272.
41. Hsu, M.H.; Tsai, M.F.; Chen, H.H. Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. *Lect. Note. Comput. Sci.* **2006**, *4182*, 1–13.
42. Hsu, M.H.; Tsai, M.F.; Chen, H.H. Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach. *Lect. Note. Comput. Sci.* **2008**, *4993*, 213–224.

43. Wang, X.; Liu, J. A Query Optimization Based on Semantic User Focus. In Proceedings of the 2009 First International Workshop on Database Technology and Applications, Hubei, China, 25–26 April 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 513–516.
44. Andreasen, T.; Bulskov, H. On Ontology-Based Querying. In Proceedings of the Flexible Query Answering Systems, Advances in Soft Computing, Warsaw, Poland, 25–27 October 2000; Physica-Verlag: Wurzburg, Germany, 2000; pp. 15–26.
45. Díaz-Galiano, M.C.; Martín-Valdivia, M.; Ure na-López, L.A. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Comput. Biol. Med.* **2009**, *39*, 396–403.
46. Carstens, C. Effects of Using a Research Context Ontology for Query Expansion. In Proceedings of the 6th European Semantic Web Conference on The Semantic Web, Crete, Greece, 31 May–4 June 2009; Springer-Verlag: Berlin, Heidelberg, Germany, 2009; pp. 919–923.
47. Cinque, L.; Malizia, A.; Navigli, R. OntoDoc: An Ontology-Based Query System for Digital Libraries. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; IEEE Computer Society: Los Alamitos, CA, USA, 2004; pp. 671–674.
48. Baziz, M.; Boughanem, M.; Pasi, G.; Prade, H. An Information Retrieval Driven by Ontology: From Query to Document Expansion. In Proceedings of the 8th International Conference on Computer-Assisted Information Retrieval, Pittsburgh, PA, USA, 30 May–1 June 2007; CID: Paris, France, 2007.
49. Liu, F.; Yu, C.; Meng, W. Personalized Web Search For Improving Retrieval Effectiveness. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 28–40.
50. Lee, M.C.; Tsai, K.H.; Wang, T.I. A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Comput. Edu.* **2008**, *50*, 1240–1257.
51. Li, X.; Chen, S. Personalized Query Expansion Based on Semantic User Model in E-learning System. In Proceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 14–16 August 2009; IEEE Press: Piscataway, NJ, USA, 2009; pp. 314–318.
52. Li, Y.; Zhong, N. Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 554–568.
53. Ravindran, D.; Gauch, S. Exploiting hierarchical relationships in conceptual search. In Proceedings of the 13th ACM International Conference on Information and Knowledge Management, Washington, D.C., USA, 8–13 November 2004; ACM: New York, NY, USA, 2004; pp. 238–239.
54. Murray, W.R. A Practical Approach to Bayesian Student Modeling. In Proceedings of the 4th International Conference on Intelligent Tutoring Systems, San Antonio, Texas, USA, 16–19 August 1998; Springer-Verlag: London, UK, 1998; pp. 424–433.
55. Fu, G.; Jones, C.B.; Abdelmoty, A.I. Ontology-Based Spatial Query Expansion in Information Retrieval. *Lect. Note. Comput. Sci.* **2005**, *3761*, 1466–1482.
56. Speretta, M.; Gauch, S. Miology: A Web Application for Organizing Personal Domain Ontologies. In Proceedings of the 2009 International Conference on Information, Process, and Knowledge

Management, Cancun, Quintana Roo, Mexico, 1–7 February 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 159–161.

57. Bhogal, J.; Macfarlane, A.; Smith, P. A review of ontology based query expansion. *Inform. Process. Manage.* **2007**, *43*, 866–886.
58. Efthimiadis, E.N. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 989–1003.
59. Suomela, S.; Kekäläinen, J. Ontology as a Search-Tool: A Study of Real Users' Query Formulation With and Without Conceptual Support. *Lect. Note. Comput. Sci.* **2005**, *3408*, 315–329.
60. Jimeno-Yepes, A.; Berlanga-Llavori, R.; Rebholz-Schuhmann, D. Ontology refinement for improved information retrieval. *Inform. Process. Manage.* **2010**, *46*, 426–435.
61. Sieg, A.; Mobasher, B.; Burke, R. Learning Ontology-Based User Profiles: A semantic Approach to Personalized Web Search. *IEEE Intell. Inform. Bull.* **2007**, *8*, 7–18.
62. Sieg, A.; Mobasher, B.; Burke, R. Web search personalization with ontological user profiles. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisboa, Portugal, 6–9 November 2007; ACM: New York, NY, USA, 2007; pp. 525–534.
63. Daoud, M.; Tamine-Lechani, L.; Boughanem, M.; Chebaro, B. A session based personalized search using an ontological user profile. In Proceedings of the 2009 ACM Symposium on Applied Computing, Honolulu, HI, USA, 8–12 March 2009; ACM: New York, NY, USA, 2009; pp. 1732–1736.
64. Chirita, P.A.; Nejdl, W.; Paiu, R.; Kohlschutter, C. Using ODP metadata to personalize search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 15–19 August 2005; ACM: New York, NY, USA, 2005; pp. 178–185.
65. Daoud, M.; Tamine-Lechani, L.; Boughanem, M. Learning user interests for a session-based personalized search. In Proceedings of the 2nd International Symposium on Information Interaction in Context, London, UK, 14–17 October 2008; ACM: New York, NY, USA, 2008; pp. 57–64.
66. Trajkova, J.; Gauch, S. Improving Ontology-based User Profiles. In Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval, Vaucluse, France, 26–28 April 2004; CID: Paris, France, 2004; pp. 380–389.
67. Klir, G.J.; Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1995.
68. Zadeh, L.A. Fuzzy Sets. *Inform. Control* **1965**, *8*, 338–353.
69. Santini, S. An Oddly-Positioned Position Paper on Context and Ontology. In Proceedings of the 2th IEEE International Conference on Semantic Computing, Santa Clara, CA, USA, 4–7 August 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 307–314.
70. Santini, S.; Dumitrescu, A. Context as a Non-ontological Determinant of Semantics. *Lect. Note. Comput. Sci.* **2008**, *5392*, 121–136.
71. Eriksson, H. The semantic-document approach to combining documents and ontologies. *Int. J. Man-Mach. Stud.* **2007**, *65*, 624–639.

72. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: a core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, 8–12 May 2007; ACM: New York, NY, USA, 2007; pp. 697–706.
73. Suchanek, F.M.; Kasneci, G.; Weikum, G. YAGO: A Large Ontology from Wikipedia and WordNet. *J. Web Sem.* **2008**, *6*, 203–217.
74. Ren, A.; Du, X.; Wang, P. Ontology-Based Categorization of Web Search Results Using YAGO. In Proceedings of the Second International Joint Conference on Computational Sciences and Optimization, Sanya, Hainan, China, 24–26 April 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 800–804.
75. Kasneci, G.; Suchanek, F.M.; Ifrim, G.; Ramanath, M.; Weikum, G. NAGA: Searching and Ranking Knowledge. In Proceedings of the 24th International Conference on Data Engineering, Cancún, México, 7–12 April 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 953–962.
76. Kasneci, G.; Suchanek, F.M.; Ifrim, G.; Elbassuoni, S.; Ramanath, M.; Weikum, G. NAGA: Harvesting, searching and ranking knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; ACM: New York, NY, USA, 2008; pp. 1285–1288.
77. Answers, Yahoo! Available online: http://it.answers.yahoo.com/ (accessed on 28 August 2010).
78. START. Available online: http://start.csail.mit.edu/ (accessed on 28 August 2010).
79. Dudev, M.; Elbassuoni, S.; Luxenburger, J.; Ramanath, M.; Weikum, G. Personalizing the Search for Knowledge. In Proceedings of the 2nd International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases, Auckland, New Zealand, 23 August 2008; Available online: http://persdb08.stanford.edu/home.html/ (accessed on 18 October 2010).
80. Xu, S.; Bao, S.; Fei, B.; Su, Z.; Yu, Y. Exploring folksonomy for personalized search. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore 20–24 July 2008; ACM: New York, NY, USA, 2008; pp. 155–162.
81. Kim, H.L.; Scerri, S.; Breslin, J.G.; Decker, S.; Kim, H.G. The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, Berlin, Germany, 22–28 September 2008; pp. 128–137.
82. Noll, M.G.; Meinel, C. Web Search Personalization Via Social Bookmarking and Tagging. *Lect. Note. Comput. Sci.* **2007**, *4825*, 367–380.
83. Vallet, D.; Cantador, I.; Jose, J.M. Personalizing Web Search with Folksonomy-Based User and Document Profiles. *Lect. Note. Comput. Sci.* **2010**, *5993*, 420–431.

84. Yeung, C.A.; Gibbins, N.; Shadbolt, N. A Study of User Profile Generation from Folksonomies. In Proceedings of the WWW 2008 Workshop on Social Web and Knowledge Management, Beijing, China, 21–25 April 2008.