The Knowledge Engineering Review, Vol. 23:1, 73–100. © 2008, Cambridge University Press
doi:10.1017/S0269888907001282 Printed in the United Kingdom

Personalized information retrieval based on context and ontological knowledge

PH. MYLONAS[1], D. VALLET[2], P. CASTELLS[2], M. FERNÁNDEZ[2] and Y. AVRITHIS[1]

[1]*National Technical University of Athens, Image, Video and Multimedia Laboratory, 9, Iroon Polytechniou street, 15773 Zographou, Athens, Greece;*
e-mail: fmylonas@image.ntua.gr, iavr@image.ntua.gr
[2]*Universidad Autónoma de Madrid, Escuela Politécnica Superior, 28049 Madrid, Spain;*
e-mail: david.vallet@uam.es, pablo.castells@uam.es, miriam.fernandez@uam.es

Abstract

Context modeling has long been acknowledged as a key aspect in a wide variety of problem domains. In this paper we focus on the combination of contextualization and personalization methods to improve the performance of personalized information retrieval. The key aspects in our proposed approach are (1) the explicit distinction between historic user context and live user context, (2) the use of ontology-driven representations of the domain of discourse, as a common, enriched representational ground for content meaning, user interests, and contextual conditions, enabling the definition of effective means to relate the three of them, and (3) the introduction of fuzzy representations as an instrument to properly handle the uncertainty and imprecision involved in the automatic interpretation of meanings, user attention, and user wishes. Based on a formal grounding at the representational level, we propose methods for the automatic extraction of persistent semantic user preferences, and live, ad-hoc user interests, which are combined in order to improve the accuracy and reliability of personalization for retrieval.

1 Introduction

The notion of context (McCarthy, 1993; Edmonds, 1999) has long been acknowledged as being of key importance in a wide variety of fields, such as mobile and pervasive computing (Brown *et al.*, 1997; Hong & Landay, 2001; Heer *et al.*, 2003; Coutaz *et al.*, 2005), computational linguistics (Wiebe *et al.*, 1996; van Eijck, 2000; Finkelstein *et al.*, 2002), automatic image analysis (Dasiopoulou *et al.*, 2005; Mylonas and Avrithis, 2005), or information retrieval (Bharat, 2000; Lawrence, 2000; Finkelstein *et al.*, 2002; Haveliwala, 2002), to name a few. A considerable body of research in such areas has investigated the representation and usage of context as a key element for example, to enhance the understanding of human speech, needs, activities, and intentions (Lewis, 1980), to raise the system awareness of the external conditions that may influence human priorities and plans, to build an awareness of the available resources for the system to accomplish a certain goal, and in general, to better grasp the relative nature of truth.

The research presented here focuses on the role of context in information retrieval (IR), and more specifically, in its smooth integration into the personalization of content retrieval. Personalization seeks to improve the subjective performance of retrieval as perceived by individual users (Jeh & Widom, 2003; Gauch *et al.*, 2004; Liu *et al.*, 2004; Micarelli & Sciarrone, 2004; Castells *et al.*, 2005). Our work aims at improving the effectiveness of personalization as perceived in a specific context, reducing some of its occasional drawbacks, such as obtrusiveness, inaccuracy, inconsistency, and distraction, by making it more *context-relevant* and contextually coherent.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

The models and techniques proposed here address the automatic extraction of persistent, content-based user preferences, as well as live ad-hoc user interests, in such a way that the combination of both produces contextualized user models, which are then applied to gain accuracy in the personalization of retrieval results.

The goal of enhancing IR models and methods towards context-aware models has raised increasing interest in the research community (Kim & Chan, 2003; Shen *et al.*, 2005), and is being identified as a key step in order to cope with the continuous growth of information environments (repositories, networks, users) worldwide, which may pose serious challenges to current search technologies in the future. In an increasingly demanding and competitive market, user queries alone are often not enough for a modern search engine to answer information needs in an effective way, meeting user expectations. For a complex or difficult information request, the user may need to modify his/her query and view ranked documents in many iterations before the information need is satisfied. In such an interactive retrieval scenario, the information naturally available to the retrieval system is more than just the current user query and the document collection—in general, arbitrary interaction history can be made available to the retrieval system, including past queries, the documents that the user has chosen to view, and even how a user has accessed a document.

Several context-sensitive retrieval algorithms exist indeed in the literature, most of them based on statistical language models to combine the preceding queries and clicked document summaries with the current query, for better ranking of documents (Bharat, 2000; Lawrence, 2000; Finkelstein *et al.*, 2002; Haveliwala, 2002). Relevance feedback (Rocchio, 1971), and later, implicit feedback (Campbell & Van Rijsbergen, 1966; Kelly & Teevan, 2003; Staab & Studer, 2004; White *et al.*, 2004), similarly exploit contextual-user input as a source of information to complement explicit user queries and guide the retrieval process. The proposal presented here has much in common with the directions explored in such works. Our research aims at enhancing the accuracy and effectiveness achieved in prior approaches by enriching and further elaborating on the proposed views in several ways, as follows.

First, most existing context-sensitive IR systems base their retrieval decision solely on queries, keywords, topics, and document collections. In contrast with this, we propose a full-fledged ontology-driven approach for an enhanced representation of the semantic context of information objects and user actions, in order to better interrelate user-sought meanings with available meanings in the search space, beyond what can be achieved using documents and keywords only. Second, the existing views on implicit user feedback and preferences as sources for IR context (which is often called 'personalized IR') do not make a clear, explicit difference between the live user context and the historical context. As a consequence, either the wider perspective of overall user trends, or the ability of the system to focus on temporary user priorities, are often lost. Our approach deals with both sides (persistent vs. live) of implicit user interests in different ways, seeking their reciprocal improvement while taking into account the different nature and most effective treatment that are proper to each. Finally, as a means to tackle the approximative nature and the inherent uncertainty involved in the automatic interpretation of meanings, user attention, and user wishes in a formal way, we propose the introduction of fuzzy representations, based on fuzzy theory (Zadeh, 1965; Klir & Bo, 1995), as a formal grounding for the development of our models and algorithms.

The rest of the paper is organized as follows. The next section introduces the basic assumptions and principles of our proposed approach to context modeling for personalized retrieval. The formal grounding of the approach in fuzzy relational algebra is described in detail in Section 3. Based on this, Section 4 describes the method to extract user interests from historic records of user interaction with a retrieval system; and Section 5 explains the contextualization of user preferences at retrieval time, for the contextual personalization of retrieval results. Section 6 shows a simple example illustrating how these techniques are applied in a detailed scenario. The empirical results obtained in two sets of experiments are reported in Section 7, after which some final conclusions are given.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

2 Ontology-based context for personalization and retrieval

Merriam-Webster's Dictionary defines context as '*the interrelated conditions in which something exists or occurs*'. In our research, the occurring events under consideration are the queries issued by a user in an interactive retrieval session towards the satisfaction of an information need. The surrounding conditions include (1) background long-term preferences, either explicitly manifested, or implicitly evidenced by the user in prior sessions with the retrieval system, (2) the short-term user focus, implicitly evidenced in her live clicks and queries during an ongoing session, and (3) the semantic scope (e.g. thematic area) of the information requested or accessed by the user in an ongoing retrieval session. The stress on *interrelation* in the above definition is of particular relevance to our view, and is treated explicitly, as will be shown. Before going further into the details of our approach, the motivation and development of our contextual notion and methods are grounded on a set of problems, assumptions, views and design decisions, which are stated next.

Our research considers the following retrieval setting: A set of users $\mathcal{U}$ interact with a retrieval space $\mathcal{D}$ through a retrieval interface including a search engine and browsing facilities. The latter allows the inspection of search result sets, or the direct navigation in the retrieval space, and the selection and display of information objects. The retrieval space $\mathcal{D}$ is made of information objects, typically (though not mandatorily) containing a fair amount of unstructured or semi-structured content, for example, text and multimedia objects and/or documents. The information objects are annotated with metadata, consisting of concepts, properties, and values defined according to a domain ontology $\mathcal{O}$, and stored in a knowledge base (KB). The ontology defines concept classes and semantic relations of arbitrary types between them, which are instantiated in the KB, forming a semantic network. The practical problems involved in meeting the latter conditions are the object of a large body of research on ontology construction (Staab & Studer, 2004), semantic annotation (Dill *et al.*, 2003; Kiryakov *et al.*, 2004; Popov *et al.*, 2004), semantic integration (Kiryakov *et al.*, 2004; Noy, 2004), and ontology alignment (Ehrig & Sure, 2004; Euzenat, 2004), and are not the focus of this paper.

In this setting, users have a-priori interests for different topics, subjects, and 'things'. Many or even most of these preferences may be unrelated to the retrieval corpus at hand, but we assume the existence of a subset of user interests (which we shall name P, for 'preference') having some kind of link to the corpus. The type and aspects of such links can be manifold, in particular, they may be related to external qualities of the information objects, such as their nature (e.g. encyclopedic, journalistic, scientific), purpose (acquiring knowledge, finding directions, having fun), quality, commercial value, former user experience with the objects, and so on. Besides these properties, it is well-known that a particularly relevant side of user interests (in terms of the value, generality, and amenability to formalization that can be brought by this view) can be related to the internal semantics conveyed by information objects, which is a common principle underlying mainstream research in the IR field (Salton, G. and McGill, 1983).

Following this common view, we define P as a set of *meanings* that can be found or referred to in items of $\mathcal{D}$. Beyond raw keywords and multimedia descriptors, which are commonly used as semantic-representation bricks for user needs in conventional IR, ontologies are being investigated in the field as enablers of qualitatively higher expressivity and precision in such descriptions (Gauch *et al.*, 2004; Kiryakov *et al.*, 2004; Popov *et al.*, 2004; Castells *et al.*, 2007). In our approach, user preferences P are described as a set of semantic entities that the user has interest for to varying degrees, where for this purpose, the same ontology as has been used to annotate the corpus is used. This provides a fairly precise, expressive, and unified representational grounding, in which both user interests and content meaning are represented in the same space, in which they can be conveniently compared (Castells *et al.*, 2005).

Nonetheless, whereas the common ontology-based view tends to lean towards an ideal view of the world, user interests are a typical example of magnitudes that can hardly be captured in a crisp, clear-cut sense. User preferences are relative, multidimensional, time-dependent, task-dependent, involving different degrees, which are dynamic and relative to a wide variety of

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

contextual factors. In addition, there is considerable uncertainty inherently involved in the representation and/or prediction of user inclinations within a software system. Contextual conditions are equally difficult to define and grasp in ways that are devoid of uncertainty and imprecision. Even content semantics are far too complex to be formally described in a complete or unambiguous manner (Kraft *et al.*, 1998), and needs to borrow further information from context to get a precise interpretation. The uncertainty increases considerably when the semantic descriptions are extracted by automatic means, through (text or multimedia) content analysis techniques. Finally, even when the meaning is clear, relations among real-life concepts are often a matter of degree, and one way to efficiently represent and model them is by the use of fuzzy relations. Taking all this into consideration, our approach complements the ontology-based perspective with fuzzy notions for the representation of user preferences, user context, content semantics, and relations between concepts. Our proposed methods for user profiling and personalized retrieval in context are founded in the principles of fuzzy sets and fuzzy relational algebra, taking advantage of the available techniques in that area, which are suitable to deal with problems involving fuzzy magnitudes (Zadeh, 1965; Klir & Bo, 1995).

In this frame where tolerance to imprecise descriptions is an assumed given, context modeling takes on a key role in harnessing the degree of fuzziness involved in the framework. The models developed in prior work (based on e.g. session-lived user input/feedback (Campbell & Van Rijsbergen, 1996; Rocchio, 1971; Kelly & Teevan, 2003; Kim & Chan, 2003; White *et al.*, 2004; Shen *et al.*, 2005), user preferences (Jeh & Widom, 2003; Gauch *et al.*, 2004; Liu *et al.*, 2004; Micarelli & Sciarrone, 2004; Castells *et al.*, 2007), ambient environment (Brown *et al.*, 1997; Hong & Landay, 2001; Heer *et al.*, 2003), and task situations (Coutaz *et al.*, 2005), spatial relations between objects in an image (Dasiopoulou *et al.*, 2005; Mylonas and Avrithis, 2005), linguistic relations between words in a text (Wiebe *et al.*, 1996; van Eijck, 2000; Finkelstein *et al.*, 2002), background topics (Haveliwala, 2002), etc. could be equally useful here to help handle this uncertainty. As a novel contribution, we propose an enhancement of such prior work, based on the exploitation of ontological information as a source of semantic context and/or an aid to relate different parts of the contextual scope in the retrieval process. The extra semantics (precise classification, explicit relations between concepts) supply a rich source of additional knowledge, enabling significant improvements with respect to the results that can be achieved by the use of unrelated plain keywords.

The notion of context takes on two perspectives in our framework, which are applied at profiling time and retrieval time, respectively. In both phases, the context consists of, put informally, a fuzzy region of a domain ontology, and is used to help focus or extend the system interpretation of user interests to a specific semantic area. In the profiling phase, which takes place off-line, the system detects user-preference patterns by analyzing a large set of recorded user actions and requests. The system analyzes the semantic relations to find common thematic ground for different subsets of the usage history, in a clustering-based approach, as will be shown in Section 4. The contextual notion applied here is taxonomic and of restrictive nature, and is used to reduce noise and uncertainty, by ignoring irrelevant user actions, and focusing on the most cohesive ones, from which it is safer to predict user interests. The taxonomic context refers to whatever is semantically common among a set of elements, which may refer to the common meaning of a set of concepts, or to the overall topic of a document, respectively. When using an ontological knowledge representation, as the one proposed herein, to interpret the meaning of an information object, it is this taxonomic context of a concept that provides its truly intended meaning. In other words, the true source of information is the semantic commonalities of certain concepts and not each one independently. The common meaning of concepts is thus used to best determine either their topics, or the associated user preferences to which they should be mapped.

At runtime, these principles are applied in a slightly different way. Even if the user is believed to have a persistent set of user interests, either learnt by the system in the profiling phase, or manually provided by the user, it is assumed that such interests are not static, but vary with time and

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

depend on the situation. Therefore, our model distinguishes a persistent component (which evolves at a slower pace) of a-priori user preferences, and a temporary, ad-hoc component, which is dependent on the live context within which the user engages in content-retrieval tasks. In our approach, the latter takes the form of an explicit, dynamic representation of the live semantic context as a fuzzy set of domain concepts, which is built by collecting ontology elements involved in user actions. This runtime representation of context is used in combination with the persistent user preferences in order to compute a focused, contextualized set of user interests. The computation of this set is achieved in two steps, consisting of a contextual expansion, followed by a contraction. In the first step, the initial preference and context sets are completed to form semantically coherent supersets (based on fuzzy compositions and unions), and in the contraction, a fuzzy intersection of the supersets is determined, as will be described in Section 5. Finally, the contextualized user interests are used to achieve a better, more accurate and reliable personalization of the retrieval results returned by the system in response to user queries.

3 Fuzzy context representation

The proposed context-based personalization model can be expressed in a formal manner with the use of basic elements towards semantic interpretation, such as concepts, relations between concepts and topics, that build an ontology structure. Since relations among real-life concepts are often uncertain or a matter of degree, which can be suitably modeled using fuzziness, the approach followed herein is based on a formal methodology and mathematical notation founded on fuzzy relational algebra (Zadeh, 1965; Klir & Bo, 1995). Its basic principles are summarized in the following subsections.

3.1 Mathematical notation

Given a universe $\mathcal{V}$, a crisp set S of concepts on $\mathcal{V}$ is described by a membership function $\mu_S : \mathcal{V} \rightarrow \{0, 1\}$. The crisp set S is defined as $S = \{s_i\}$, $i = 1, \ldots, N$. A *fuzzy* set F on S is described by a membership function $\mu_F : S \rightarrow [0, 1]$. We may describe the fuzzy set F using the well-known sum notation for fuzzy sets (Miyamoto, 1990) as:

$$F = \sum_i s_i / w_i = \{s_1/w_1, s_2/w_2, \ldots, s_n/w_n\} \tag{1}$$

where:

- $i \in N_n$, $n = |S|$ is the cardinality of the crisp set S,
- $w_i = \mu_F(s_i)$ or, more simply $w_i = F(s_i)$, is the membership degree of concept $s_i \in S$.

Consequently, equation (1) for a concept $s \in S$ can be written equivalently as:

$$F = \sum_{s \in S} s / \mu_F(s) = \sum_{s \in S} s / F(s) \tag{2}$$

The *height* of the fuzzy set F is defined as the maximum membership degree:

$$h(F) = \max_i (F(s_i)), \quad i \in N_n \tag{3}$$

A *normal* fuzzy set is defined as a fuzzy set having height $= 1$, whereas *cp* is an involutive fuzzy complement, that is, a fuzzy complement for which: $cp(cp(\alpha)) = \alpha$, for each $\alpha \in [0, 1]$ (Klir & Bo, 1995). The product of a fuzzy set F with a number $\gamma \in [0, 1]$ is defined as $[F \cdot \gamma]_{(x)} = F(x) \cdot \gamma$, $\forall x \in S$, $\gamma \in [0, 1]$.

Now, let R be the crisp set of fuzzy relations defined as:

$$R = \{R_i\},\ R_i : S \times S \rightarrow [0,\ 1], \quad i = 1, \ldots, M \tag{4}$$

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

and Z be the crisp set of concepts that at the same time are considered to be thematic topics. Then the proposed fuzzy ontology contains concepts, relations, and topics and can be formalized as follows:

$$\mathcal{O} = \{S, R, Z\} \tag{5}$$

In equation (5), $\mathcal{O}$ is a fuzzy ontology, S is the crisp set of concepts described by the ontology, R is the crisp set of fuzzy semantic relations amongst these concepts and Z is the crisp set of topics available in $\mathcal{O}$, where $Z \subset S$.

Given the set of all fuzzy sets on S, $\mathcal{F}_S$, then $F \in \mathcal{F}_S$. Let $\mathcal{U}$ be the set of all users $\hat{u}$ in our personalization framework, that is, a user $\hat{u} \in \mathcal{U}$. Let $\mathcal{P}$ be the set of all user preferences and $\mathcal{P}_\mathcal{O}$ be the set of all user preferences on $\mathcal{O}$. Then $\mathcal{P}_\mathcal{O} \subset \mathcal{F}_S$ and $\mathcal{P}_\mathcal{O} = \mathcal{F}_Z \subset \mathcal{F}_S$, whereas $P_{\hat{u}} \in \mathcal{P}_\mathcal{O}$ depicts a specific user preference and is described as a fuzzy set on Z. Since the fact that a user preference is relative to a user is clear, in the following we shall omit $\hat{u}$ as the index variable and use just P for short, as long as the meaning is clear.

Furthermore, let $\mathcal{C}_\mathcal{O}$ denote the set of all contexts on $\mathcal{O}$, $\mathcal{C}_\mathcal{O} \subseteq \mathcal{F}_S$. Similar to the user preferences case, $\hat{C} \in \mathcal{C}_\mathcal{O}$ is a fuzzy set on the crisp set of concepts S and symbolizes the runtime context and let C denote the set of all runtime contexts. Let us also denote the crisp set of concepts characterizing the crisp (taxonomic) context as C', whereas its fuzzy counterpart C provides the taxonomic context in the form of a fuzzy set of concepts on S, $C \in \mathcal{C}_\mathcal{O}$. Finally, let $\mathcal{D}$ be the crisp set of all available information objects (e.g. text or multimedia documents), S_d the fuzzy set of concepts associated to $d \in \mathcal{D}$, where $S_d \in \mathcal{F}_S$, and $I(s, d)$ the constructed semantic index between documents and concepts (Al-Khatib *et al.*, 1990).

As the last step, we define the contextualization of user preferences as a mapping $\Phi : \mathcal{P} \times \hat{C} \rightarrow \mathcal{P}$ so that for all $p \in \mathcal{P}$ and $c \in \hat{C}$, $p \models \Phi(p, c)$. In this context the entailment $p \models q$ means that any consequence that could be inferred from q could also be inferred from p. For instance, given a user $\hat{u} \in \mathcal{U}$, if $P_{\hat{u}} = q$ implies that $\hat{u}$ 'likes x' (whatever this means), then $\hat{u}$ would also 'like x' if his/her preference was p.

3.2 Fuzzy semantic relations

In order to define, extract, and use both the taxonomic and runtime contexts of a set of concepts, we rely on the semantics of their fuzzy semantic relations. As discussed in the previous subsection, a *fuzzy binary relation* on S is defined as a function $R_i : S \times S \rightarrow [0, 1], i = 1, \ldots, M$. The inverse relation of relation $R_i(x, y)$, $x, y \in S$ is defined as $R_i^{-1}(x, y) = R_i(y, x)$. We use the prefix notation $R_i(x, y)$ for fuzzy relations, rather than the infix notation xR_iy, since the former is considered to be more convenient for the reader. The *intersection*, *union*, and sup-*t composition* of any two fuzzy relations R_1 and R_2 defined on the same set of concepts S are given by:

$$(R_1 \cap R_2)(x, y) = t(R_1(x, y), R_2(x, y)) \tag{6}$$

$$(R_1 \cup R_2)(x, y) = u(R_1(x, y), R_2(x, y)) \tag{7}$$

$$(R_1 \circ R_2)(x, y) = \sup_{w \in S} t(R_1(x, w), R_2(w, y)) \tag{8}$$

where t and u are a fuzzy t-norm and a fuzzy t-conorm, respectively. The standard t-norm and t-conorm are the *min* and *max* functions, respectively, but others may be used if appropriate. The operation of the union of fuzzy relations can be generalized to M relations. If $R_1, R_2, \ldots, R_M$ are fuzzy relations in $S \times S$ then their union R^u is a relation defined in $S \times S$ such that for all $(x, y) \in S \times S$, $R^u(x, y) = u(R_i(x, y))$. A transitive closure of a relation R_i is the smallest transitive relation that contains the original relation and has the fewest possible members. In general,

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

the closure of a relation is the smallest extension of the relation that has a certain specific property such as the reflexivity, symmetry, or transitivity, as the latter is defined in (Klir & Bo, 1995). The sup-t transitive closure $Tr^t(R_i)$ of a fuzzy relation R_i is formally given by:

$$Tr^t(R_i) = \bigcup_{j=1}^{\infty} R_i^{(j)} \tag{9}$$

where $R_i^{(j)} = R_i \circ R_i^{(j-1)}$ and $R_i^{(1)} = R_i$. It is proved that if R_i is reflexive, then its transitive closure is given by $Tr^t(R_i) = R_i^{(n-1)}$, where $n = |S|$ (Klir & Bo, 1995).

Based on the relations R_i we first construct the following combined relation T utilized in the definition of the taxonomic context C:

$$T = Tr^t(\bigcup_i R_i^{p_i}), \quad p_i \in \{-1, 0, 1\}, \quad i = 1, \ldots, M \tag{10}$$

where the value of p_i is determined by the semantics of each relation R_i used in the construction of T. More specifically:

- $p_i = 1$, if the semantics of R_i imply it should be considered as is
- $p_i = -1$, if the semantics of R_i imply its inverse should be considered
- $p_i = 0$, if the semantics of R_i do not allow its participation in the construction of the combined relation T.

The transitive closure in equation (10) is required in order for T to be taxonomic, as the union of transitive relations is not necessarily transitive, independent of the fuzzy t-conorm used. In the above context, a fuzzy semantic relation defines, for each element $s \in S$, the fuzzy set of its ancestors and its descendants. For instance, if our knowledge states that 'American Civil War' is before 'WWI' and 'WWI' is before 'WWII', it is not certain that it also states that 'American Civil War' is before 'WWII'. A transitive closure would correct this inconsistency. Similarly, by performing the respective closures on relations that correlate a pair of concepts of the same set, we enforce their consistency.

Similarly, based on a different subset of relations R_i, we construct the combined relation $\hat{T}$ for use in the determination of the runtime context $\hat{C}$:

$$\hat{T} = \bigcup_i (R_i^{\hat{p}_i}), \quad \hat{p}_i \in \{0, 1\}, \quad i = 1, \ldots, \hat{M} \tag{11}$$

For the purpose of analyzing text and multimedia document descriptions, relation T has been generated with the use of a small set of fuzzy taxonomic relations, whose semantics are derived from the MPEG-7 standard (ISO/IEC FDIS 15938-5, 2001) and are summarized in Table 1. On the other hand, relation $\hat{T}$ has been constructed with the use of the entire set of relations available in the knowledge base. This approach is ideal for the interpretation of the two kinds of context and user preferences followed herein; initially, when dealing with the generic user profile, focus is given on the semantics of high-level abstract concepts, whereas during the retrieval phase, additional precision and a more specific view is required as the runtime preference expansion takes place. The latter demands the use of all available information in the KB. Of course, as the construction of relation $\hat{T}$ implies, an intermediate step of removing its possible cycles, that are present due to the utilization of all relations and their inverses, is necessary before the application of the taxonomy-based expansion process. This step is analyzed in detail further in Subsection 5.1 of the current manuscript.

The aforementioned relations are traditionally defined as crisp relations. However, in this work we consider them to be fuzzy, where fuzziness has the following meaning: high values of $Sp(a, b)$, for instance, imply that the meaning of b approaches the meaning of a, while as $Sp(a, b)$ decreases, the meaning of b becomes narrower than the meaning of a. A similar meaning is given to fuzziness of the rest of the semantic relations of Table 1, as well. Based on the fuzzy roles and semantic

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Table 1 Taxonomic relations used for generation of combined relation T

Name	Inverse	Symbol	Meaning	Example	
				a	b
Specialization	Generalization	$Sp(a,b)$	b is a specialization in the meaning of a	mammal	dog
Part	PartOf	$P(a,b)$	b is a part of a	London	Soho
Example	ExampleOf	$Ex(a,b)$	b is an example of a	president	Clinton
Instrument	InstrumentOf	$Ins(a,b)$	b is an instrument of or is employed by a	cut	knife
Location	LocationOf	$Loc(a,b)$	b is the location of a	concert	stage
Patient	PatientOf	$Pat(a,b)$	b is affected by or undergoes the action of a	give	book
Property	PropertyOf	$Pr(a,b)$	b is a property of a	banana	ripeness

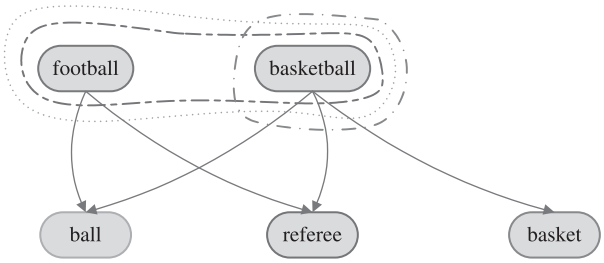


Figure 1 Taxonomic context example

interpretations of R_i, it is easy to see that both the aforementioned relations (10) and (11), combine them in a straightforward and meaningful way, utilizing inverse functionality where it is semantically appropriate. More specifically, relation T utilizes the following subset of relations:

$$T = Tr^t(Sp \cup P^{-1} \cup Ex \cup Ins \cup Loc^{-1} \cup Pat \cup Pr^{-1}) \tag{12}$$

Relation T is of great importance, as it allows us to define, extract, and use the taxonomic context of a set of concepts. All relations used for its generation are partial taxonomic relations, thus abandoning properties like synonymity. Still, this does not entail that their union is also antisymmetric. Quite the contrary, T may vary from being a partial taxonomic to being an equivalence relation. This is an important observation, as true semantic relations also fit in this range (total symmetricity, as well as total antisymmetricity often have to be abandoned when modeling real-life relationships). Still, the taxonomic assumption and the semantics of the used individual relations, as well as our experiments, indicate that T is 'almost' antisymmetric and we may refer to it as ('almost') taxonomic. Relying on its semantics, we define the crisp taxonomic context $\hat{C}$ of a single concept $s \in S$ as the set of its antecedents provided by relation T in the ontology.

As observed in Figure 1, concepts *football* and *basketball* are the antecedents of concepts *ball* and *referee* in relation T, whereas concept *basketball* is the only antecedent of concept *basket*. More formally, following the standard superset/subset notation from fuzzy relational algebra, the crisp context $C'(s)$ of a single concept $s \in S$ is given by:

$$C'(s) = T_{\leq}(s) \tag{13}$$

Assuming again that a set of concepts S is crisp, that is, that all considered concepts belong to the set with degree one, the context of the entire set, which is again a set of concepts, can be defined simply as the set of their common antecedents:

$$C'(S) = \cap C'(s_i), \quad s_i \in S \tag{14}$$

As represented in Figure 1, concept *basketball* is the only common antecedent of all three concepts *ball*, *referee*, and *basket* in relation T, that is, *basketball* is the context of *ball*, *referee*, and *basket*.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

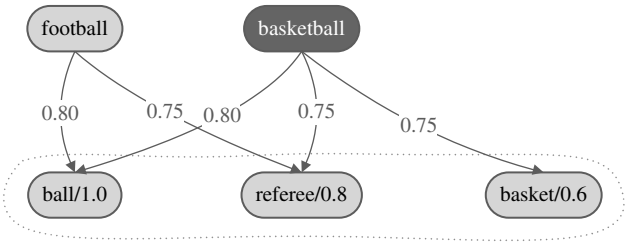


Figure 2 Fuzzy taxonomic context example

As more concepts are considered, the context becomes narrower, that is, it contains less concepts and to smaller degrees. When the definition of context is extended to the case of fuzzy sets of concepts (Figure 2), the crisp taxonomic context C' is replaced by its fuzzy counterpart, that is, the fuzzy taxonomic context C. Obviously, the semantic meaning of fuzzy context remains the same as in the crisp case, that is, the above property must still hold. The context C of the normal fuzzy set F on S is calculated as:

$$C(F) = \bigcap_i \mathcal{K}(s_i),\ s_i \in F \tag{15}$$

where $\mathcal{K}(s_i)$ is the 'considered' context of s_i, that is, the concept's context when taking its degree of participation to the set into account. $\mathcal{K}(s_i)$ is defined as:

$$\mathcal{K}(s_i) \doteq C'(s_i) \cup cp(S \cdot F(s_i)) \tag{16}$$

where $S{\cdot}F(s_i)$ is the product of set S with the membership degree $F(s_i)$ as defined in Subsection 3.1, the '$\doteq$' sign designates equality that comes from definition, cp is a fuzzy involutive complement and $C'(s_i)$ denotes the crisp context of a single concept s_i.

Moreover, we observe that because of the nature of fuzzy sets the following properties hold as well:

- $F(s_i) = 0 \Rightarrow C(F) = C(F - \{s_i\})$, that is, no context narrowing
- $F(s_i) = 1 \Rightarrow C(F) \subseteq C(s_i)$, that is, full narrowing of context
- $C(F)$ decreases monotonically with respect to $F(s_i)$.

Considering the semantics of the T relation and the above process of context determination, it is easy to realize that when the concepts in a set are highly related to a common meaning, the context will have high degrees of membership for the concepts that represent this common meaning. Therefore, the height of the context $h(C(F))$ will be used in the following as a measure of the semantic correlation of concepts in set F. This measure represents also the degree of relevance of the concepts in the set.

4 Profiling

So far, we illustrated the modeling of contextual dependence between concepts and relations using a fuzzy algebra representation and two constructed semantic relations. We continue with presenting the role of user profiling in our personalization approach, the notion of user preferences, as well as the presentation, extraction, and use of these preferences in the process.

4.1 The role of user profiles

It is a fact that uncertainty is inherent to the process of information retrieval (Chang & Hsu, 1998; Kraft *et al.*, 1998), as a limited set of terms cannot fully describe the user's wish. The role of

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

personalization is to reduce this uncertainty, by using more information about the user's wishes than just the local interest. The contribution of *user profiles* in understanding the effect inherent in information retrieval, when two distinct users presenting identical queries obtain different subsets of retrieved documents and to different degrees, is crucial. The user profile is generated through the constant monitoring of the user's interaction, which contains less uncertainty because of the nature of his/her actions, as long as the monitoring period is sufficient and representative of the user's preferences. Therefore, a user profile, which contains valuable information concerning the user's global interest, that is, information concerning the user's preferences over a long period of time, may be used whenever the query, that is, the user's local preference or in other words the scope of his/her current interaction, provides insufficient information about the user and his/her local interest.

In order to process the user profile using the stored knowledge, the representation of the former needs to be compatible with the underlying ontological knowledge. As all the relations R_i that exist in the ontology $\mathcal{O}$ are defined on the crisp set S of concepts, we defined user preferences on the same set: their representation, which also allows for degrees of preference, is the usage of a single fuzzy set defined on the set of concepts, as described in Subsection 3.1.

When the user poses a query that is in fact related to one of his/her preferences, that preference may be used to facilitate the interpretation of the query, as well as the ranking of the selected documents. However, usage of preferences that are unrelated to the query may only be viewed as addition of noise, as any proximity between selected documents and these preferences is coincidental in the given context. Thus, in addition to positive preferences P^+, special care must be taken for the representation and separate store of negative preferences P^-, so that they are processed separately. Consequently, a user preference P is actually constituted by two distinct fuzzy sets of preferences:

$$P = \{P^+, P^-\}, \quad P \in \mathcal{F}_z \tag{17}$$

4.2 User actions

In the process of identifying both kinds of user preferences, we start from the set of documents available in each user's usage history. The proposed user-profile implementation receives this usage history as input and produces the corresponding user preferences as output. In order to achieve this, the process also accesses the semantic index and the ontology. The set of documents available in the usage history is constructed as a result of the application of four user action types, during the user's interaction with the IR system. These actions characterize the user and express his/her personal view of the search space content. These actions are directly associated to user requests or queries and, therefore, we shall use the term *query* in the following. The four possible content-retrieval user action types that a user may pose as queries in our framework are: *keyword-based* queries, *view document* queries, *relevance-feedback* queries, and *browsing* queries. We define the set of each query type as a fuzzy set of concepts on S, whose degrees of membership are obtained by monitoring the specific query type appearance probabilities, as follows:

1. Keyword-based queries Q^k, formally defined as a fuzzy set of concepts on S, that is, $Q^k \in \mathcal{F}_S$. Keywords may be extracted from a natural language or a keyword-based encountered query and are mapped to concepts in the annotation of documents, utilizing state-of-the-art information extraction techniques (Popov *et al.*, 2004).
2. View document queries Q^v, formally defined as a fuzzy set of concepts on S, that is, $Q^v \in \mathcal{F}_S$. In this case, concepts are directly encountered in the annotation of one document and retrieved with the help of the semantic index (Vallet *et al.*, 2005), consequently $Q^v = S_d$ in this case.
3. Relevance-feedback queries Q^{rf}, satisfying users' relevance-feedback requests and consisting of two parts, namely positive and negative relevance-feedback requests:

$$Q^{rf} = \{Q^+, Q^-\} \tag{18}$$

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Both requests are again defined as fuzzy sets on the set of concepts S, that is, $Q^+, Q^- \in \mathcal{F}_S$. Q^+ corresponds to the annotation of the set of documents marked as relevant by the user and therefore is defined as:

$$Q^+ = \underset{d \in D^+}{\cup} S_d \tag{19}$$

whereas Q^- corresponds to the annotation of the set of documents marked as non-relevant and therefore is defined as:

$$Q^- = \underset{d \in D^-}{\cup} S_d \tag{20}$$

D^+ denotes the set of documents indicated as positive by the user during the relevance feedback iterations, while D^- denotes the set of documents indicated as negative at the same time.

4. Browsing queries Q^b, according to one specific browsing topic or category of documents or concepts. Q^b is defined as a fuzzy set of topics requested for browsing by the user, that is, it is formally identified as $Q^b \in \mathcal{F}_z$.

4.3 Usage history

The user's usage history comprises a combination of all types of actions, provided that a user is able to perform any type of action at a given time. An association between the related history documents and concepts exists through the utilization of the semantic index, which is a priori constructed during analysis of either the raw content, or the associated textual annotation. Let us formally denote the entire history of each user, that is, the concepts associated to his/her usage of history documents by:

$$H = \{H^+, H^-\} \tag{21}$$

consisting of both positive H^+ and negative H^- parts. It should have been clear by now, that the term *positive* corresponds to the user's likes, whereas the term *negative* corresponds to the user's dislikes. In this context, H^+ is defined as the fuzzy set of concepts obtained by the union of all concepts related to all queries of the user, thus:

$$H^+ = Q^k \cup Q^v \cup Q^+ \cup Q^b, \quad H^+ \in \mathcal{F}_S \tag{22}$$

and

$$H^- = Q^-, \quad H^- \in \mathcal{F}_S \tag{23}$$

4.4 From documents to user preferences

4.4.1 Overview

The formal definition of user preferences as a fuzzy set of concepts described in Subsection 4.1, allows participation of a single concept in multiple preferences and to different degrees. As already stated, the history H of the user is represented as a fuzzy set on the set of concepts that are related to it and consists of both positive and negative parts. Preferences are mined by using both of these parts as input and by applying clustering algorithms on them. Utilizing the notion of context in the process, we finally extract two distinct sets of positive and negative user preferences as output and combine them in a meaningful way to obtain P.

Most clustering methods found in the literature belong to either of two general categories, partitioning and hierarchical (Theodoridis & Koutroumbas, 1998). Hierarchical methods do not require the number of clusters as input, in contrast to their partitioning counterparts. Since the number of preferences that may be encountered in a document is not known beforehand, the latter is inapplicable (Miyamoto, 1990). The same applies to the use of a supervised clustering method which allows one concept to belong to two or more clusters, such as fuzzy c-means (Benkhalifa *et al.*, 1999). This algorithm requires the number of concept clusters as input, that

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

is, it uses a hard termination criterion on the amount of clusters and thus cannot be adopted to the problem at hand. Instead, we use a hybrid approach, based on fuzzification of an agglomerative[1] hierarchical-clustering algorithm.

Letting $K' = \{k_i'\}$ be the set of crisp clusters detected in H^+, each cluster k_i' is a crisp set of concepts. However, this is not sufficient for our approach, as we need to support documents belonging to multiple distinct preferences by different degrees and at the same time retain the robustness and efficiency of the hierarchical-clustering approach. Thus, without any loss of functionality or increase of computational cost we replace the crisp clusters k_i' with fuzzy normalized clusters k_i, by constructing a fuzzy classifier from $K' = \{k_i'\} \rightarrow K = \{k_i\}$, where $K = \{k_i\}$ is the set of the obtained fuzzy clusters of concepts. As described in the following, for each fuzzy cluster k_i we obtain the fuzzy set of preferences associated to it, by exploiting its context and cardinality information. Then, by aggregating the process to the entire set of fuzzy clusters, we identify the fuzzy set of preferences related to the initial set of documents in the user's usage history, after limiting it according to the predefined set of all possible user preferences.

The sections below provide details on the initial concept-clustering process, the cluster *fuzzification*, as well as the final user preference extraction. This threefold model can be formalized in an abstract way as a function

$$\mathcal{Y} = G(\mathcal{X}) \tag{24}$$

without any assumption on how the input or output of the function may be represented and instantiated. The function takes a fuzzy set $\mathcal{X}$ as input and provides a different fuzzy set $\mathcal{Y}$ as its output. In this context, we may particularize the above statement for the specific case of positive user preferences and usage history; function G can be utilized to obtain P^+ from H^+ as:

$$P^+ = G(H^+) \tag{25}$$

The proposed approach may then be decomposed into the following four general steps:

1. Perform a crisp clustering of concepts H^+ in order to determine the count of distinct positive preferences P^+ that a history document is related to.
2. Construct a fuzzy classifier that measures the degree of correlation of a concept s_j with cluster k_i'.
3. Consider the context and cluster cardinality of the resultant fuzzy clusters k_i and mathematically adjust their computed values so as to match their semantically anticipated counterparts.
4. Identify the positive user preferences P^+ that are related to each cluster, according to the a priori known set of all possible user preferences, in order to acquire an overall result.

The same applies in the case of the application of function G to H^-, in order to obtain P^- as:

$$P^- = G(H^-) \tag{26}$$

As already stated, the final set of preferences P that correspond to the user's history is the set of positive P^+ meaningfully combined with the set of negative preferences P^-. Using again the sum notation for fuzzy sets, this may be represented as:

$$P = \sum_{s \in S} s/max(0, P^+(s) - P^-(s)) \tag{27}$$

where $P(s) = max(0, P^+(s) - P^-(s))$ denotes the final preference membership degree for each concept s.

[1] Hierarchical methods are divided into agglomerative and divisive. The former are more widely studied and applied, as well as more robust and therefore are followed herein.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

4.4.2 *Crisp clustering*

The first step towards identification of user preferences is the implementation of crisp clustering on the set of concepts that exist in the usage history. The general structure of a hierarchical-clustering approach, adjusted for the needs of the problem at hand, is as follows. Without loss of generality, we particularize our approach for positive preferences P^+, keeping in mind that the same applies for negative ones P^-.

1. When considering the available set of concepts to be clustered H^+, turn each one of them into a singleton, that is, into a cluster $k_i^{'}$ of its own.
2. For each pair of clusters $k_1^{'}$, $k_2^{'}$ calculate their compatibility indicator $d(k_1^{'}, k_2^{'})$. The $d(k_1^{'}, k_2^{'})$ is also referred to as cluster similarity, or distance metric (Kohavi & Sommerfield, 1995; Mylonas *et al.*, 2004).
3. Merge the pair of clusters that have the best compatibility indicator $d(k_1^{'}, k_2^{'})$. Depending on whether this is a similarity or a distance metric, the best indicator could be selected using the *max* or *min* operator, respectively.
4. Continue at step 2, unless termination criteria are met; termination criterion most commonly used is a meaningfully derived threshold for the value of the best compatibility indicator $d(k_1^{'}, k_2^{'})$.

As in all typical hierarchical-clustering approaches, the two key points in the above process are the identification of the clusters to merge at each step and the identification of the optimal terminating step. In this work, the height of the context $h(C(k_1^{'} \cup k_2^{'}))$ is used as a distance metric for two clusters $k_1^{'}$, $k_2^{'}$ quantifying their semantic correlation, as defined at the end of Subsection 3.2. The process terminates when the concepts are clustered into sets that correspond to distinct preferences, identified by the fact that their common context has low height. Therefore, the termination criterion is a threshold on the selected compatibility metric. The output of this step is a crisp set of clusters K', where each cluster $k_i^{'} \in K^{'}$ is a crisp set of concepts, $k_i^{'} \in S_d$.

4.4.3 *Cluster fuzzification*

The above clustering method determines successfully the count of distinct clusters that exist, but it only creates crisp clusters, that is, it does not allow for degrees of membership in the output and it does not allow for overlapping among the detected clusters. However, a concept may be related to a user preference to a degree other than 1 or 0 in real-life and may also be related to more than one distinct preferences. In order to overcome such problems, *fuzzification* of the clusters is needed. In particular, we construct a fuzzy classifier, that is, a function

$$G_c : S \rightarrow [0, 1] \tag{28}$$

that measures the degree of correlation of a concept $s_j \in S$ with cluster $k_i^{'}$. Apparently, a concept s_j should be considered correlated with cluster $k_i^{'}$, if it is related to the common meaning of the concepts in $k_i^{'}$. Therefore, the quantity

$$G_c(s_j) = \frac{h(C(k_i^{'} \cup \{s_j\}))}{h(C(k_i^{'}))} \tag{29}$$

forms an appropriate measure of correlation. It is easy to see that this measure has the following properties:

- $G_c(s_j) = 1$, if the semantics of s_j imply it should belong to $k_i^{'}$. For example: $G_c(s_j) = 1$, $\forall s_j \in k_i^{'}$.
- $G_c(s_j) = 0$, if the semantics of s_j imply it should not belong to $k_i^{'}$.
- $G_c(s_j) \in (0, 1)$, if s_j is neither totally related, nor totally unrelated to $k_i^{'}$.

Using this classifier, we expand the detected crisp clusters to include more concepts. Thus, cluster $k_i^{'}$ is replaced by the fuzzy cluster $k_i \supseteq k_i^{'}$ and $k_i = \sum_{s_j \in S_d} s_j / G_c(s_j)$, using again the sum notation for fuzzy sets.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

The last point to consider during the fuzzification step is the fact that, so far, the process of fuzzy hierarchical clustering has been based on the crisp set S_d, thus ignoring the fuzziness that exists in the semantic index. In order to incorporate this when calculating the clusters, we need to adjust their degrees of membership $k_i(s_j)$, according to the information present in the semantic index $I(s_j, d)$. Then each one of the resulting clusters corresponds to one of the distinct user preferences of the document. In order to determine the preferences that are related to a cluster k_i, we need to consider both its scalar cardinality $|k_i|$ and its context. Since taxonomic context has been defined only for *normal* fuzzy sets, each degree of membership is finally obtained as:

$$k_i(s_j) = \frac{t(k_i(s_j), I(s_j, d))}{h(t(k_i(s_j), I(s_j, d)))}, \ \forall s_j \in H^+ \tag{30}$$

where, due to the semantic nature of the above operation, t is an Archimedean t-norm.

4.4.4 *Fuzzy preferences extraction*

In order to identify the fuzzy set $\{W(k_i)\}$ of preferences related to the set of concepts under consideration, we need to calculate each $W(k_i)$, that is, the set of preferences related to each cluster k_i. The latter is computed as follows:

$$W(k_i) = \tilde{w}(C(k_i)) \cdot L(|k_i|) \tag{31}$$

where $\tilde{w}$ is a weak modifier and $L(\cdot)$ is a 'large' fuzzy number. The weak modifier is used in this work to adjust mathematically computed values so as to match its semantically anticipated counterparts; $\tilde{w}(a) = \sqrt{a}$ is a commonly used weak modifier (Klir & Bo, 1995). The 'large' fuzzy number models 'high cardinality' of clusters and forms a function from the set of real positive numbers to the [0,1] interval, quantifying the notion of 'large' or 'high'. Herein, the 'large' fuzzy number is defined as the triangular fuzzy number $(1.3, 3, \infty)$ (Klir & Bo, 1995)[2].

Obviously, if there is only a unique cluster k_i, then $\{W(k_i)\} = \tilde{w}(C(k_i))$ is a meaningful approach that denotes the output of the process in case of neglecting cluster cardinalities. On the other hand, when more than one cluster is detected, then it is imperative that cluster cardinalities are considered as well. Clusters of extremely low cardinality probably only contain misleading concepts, and therefore need to be ignored in the estimation of $\{W(k_i)\}$. On the contrary, clusters of high cardinality almost certainly correspond to distinct preferences and need to be considered in its estimation, according to equation (31).

The set of preferences that correspond to the set of history documents associated to the user queries is the set of preferences that belong to any of the detected clusters of concepts that index the given documents. For instance, for the set of positive preferences we have:

$$\{W(k_i)\} = \underset{k \in K}{\cup} W(k) \tag{32}$$

where $\cup$ is a fuzzy co-norm and K contains the set of clusters that have been detected in H^+. It is easy to see that $\{W(k_i)\}(s_j)$ will be high if a cluster k_i, whose context contains s_j, is detected in H^+, and additionally, if the cardinality of k_i is high and the degree of membership of s_j in the context of the cluster is also high (i.e. if the topic is related to the cluster and the cluster does not consist of misleading concepts). Finally, in order to validate the results of fuzzy classification, that is, assure that the set of topics $\{W(k_i)\}$ that correspond to the set of documents H^+ are derived from the set of all possible user preferences Z, we compute the quantity

$$P^+ = \{W(k_i)\} \cap Z \tag{33}$$

Following the exact same process for the negative preferences P^- and according to equation (27), the overall user preferences P are identified. Finally, an illustrative example is given in Figure 3. As observed in the figure, $W(k_1)$ corresponds to the set of preferences related to cluster k_i and $W(k_2)$ is the set of preferences related to cluster k_2. The set of preferences that belong to any of the two

[2] Let $a, b, c \in R$, $a < b < c$. The fuzzy number $u : R \rightarrow [0,1]$ denoted by (a, b, c) and defined by $u(x) = 0$ if $x \leq a$ or $x \geq c$, $u(x) = \frac{x-a}{b-a}$, if $x \in [a, b]$ and $u(x) = \frac{c-x}{c-b}$ if $x \in [b, c]$ is called a triangular fuzzy number.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

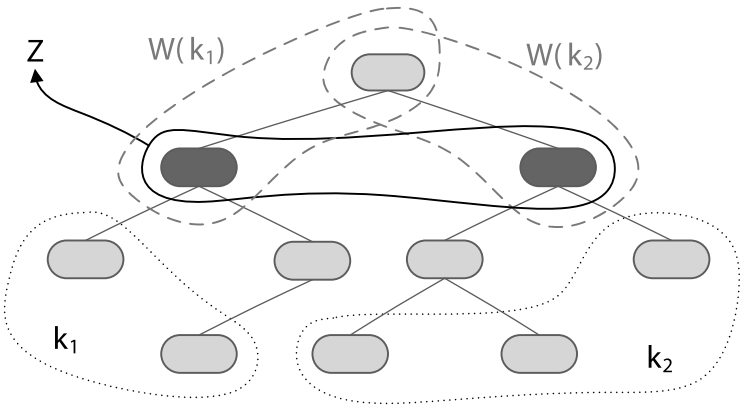


Figure 3 Relation T and fuzzy preferences extraction example

clusters is given by $W(k_1) \cup W(k_2)$, that is, the set of all three concepts. Application of equation (33) limits the set of user preferences to the two shaded topics indicated in the figure.

5 Retrieval

5.1 Contextualization

In the frame of a content-retrieval system and as already mentioned earlier, we define the semantic-retrieval runtime user context as the set of concepts that have been involved, directly or indirectly, in the interaction of a user $\hat{u}$ with the system during a retrieval session. Therefore, at each point t in time, we represent the retrieval context $\hat{C}_t(\hat{u})$ as a fuzzy set of concepts. Time is measured by the number of user requests within a session. Since the fact that the context is relative to a user is clear, in the following we shall omit this variable and use $\hat{C}_t$ as long as the meaning remains obvious.

In our approach, the semantic-runtime context $\hat{C}_t$ is built as a cumulative combination of the concepts involved in successive user requests or queries, in such a way that the importance of concepts fades away with time. This simulates the natural drift of user focus over time. Let us define the set of all available time slots as $\mathcal{T} = \{1, \ldots, M\}$, that is, $t \in \mathcal{T}$. Let us also define Q_t as the fuzzy set of concepts that is created right after each user's query at a given time t, that is, $Q_t \in \mathcal{F}_s$. Obviously, from the analysis presented in Section 4.2, we have:

$$Q_t = Q_t^k \cup Q_t^v \cup Q_t^+ \cup Q_t^b \tag{34}$$

where:

$$Q^k = \underset{t\in\mathcal{T}}{\cup} Q_t^k, \quad Q^v = \underset{t\in\mathcal{T}}{\cup} Q_t^v, \quad Q^+ = \underset{t\in\mathcal{T}}{\cup} Q_t^+, \quad Q^b = \underset{t\in\mathcal{T}}{\cup} Q_t^b \tag{35}$$

Next, the runtime context $\hat{C}_t$ at query time t is defined by combining the newly constructed fuzzy set Q_t with the runtime context $\hat{C}_{t-1}$ computed in the previous step, where the context weights computed in step $t-1$ are automatically reduced by a decay factor $\beta \in [0, 1]$. Consequently, at any given time $t > 1$, we update $\hat{C}_t$ as:

$$\hat{C}_t = \beta \hat{C}_{t-1} + (1-\beta) Q_t \tag{36}$$

where the algebraic sum and algebraic product are used for the implementation of addition and multiplication between any two given fuzzy sets (Miyamoto, 1990). Obviously $\hat{C}_t \in C_O$ and equation (36) holds for $\hat{C}_0 = \emptyset$ and $\hat{C}_1 = Q_1$. Q_t consists of a variety of user requests and the runtime context fuzzy set is not used to reformulate the query, but to focus on the preference set, thus

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

differentiating our approach from classical relevance feedback strategies (Campbell & Van Rijsbergen, 1996; Kelly & Teevan, 2003).

At this point, we have identified both the off-line representation of user preferences P associated to the set of each user's usage history H, and the runtime context $\hat{C}_t$. The selective activation of user preferences is based on finding semantic paths between preference and context concepts. The paths utilize the constructed semantic relation $\hat{T}$ between the set of concepts S available in the domain ontology $\mathcal{O}$, as described in Subsection 3.2. Our strategy consists of a semantic extension through a fuzzy semantic intersection between user preferences P and the semantic-runtime context $\hat{C}_t$.

Let us first define the notion of semantic extension of a generalized entity X with a function E. The entity X may be either the user preferences or the runtime context, as the proposed methodology is appropriate for both of them. Let us define X as a fuzzy set of concepts on S, that is, $X \in \mathcal{F}_S$. Then, similarly to the formality introduced in (9):

$$X_0 = X \quad and \quad X_{i+1} = X_i \circ \hat{T}, i > 0 \tag{37}$$

Consequently:

$$E(X) = X_{\mathcal{L}}, \tag{38}$$

where the point at which the iteration stops and equation (37) converges is denoted by $\mathcal{L}$. The iteration stops when the result from the previous iteration step is equal to the result of the current iteration step, that is, $X_{\mathcal{L}} = X_{\mathcal{L}-1}$, or in other words when $X_{\mathcal{L}-1} \circ \hat{T} = X_{\mathcal{L}-1}$. Note that in general the graph defined by $\hat{T}$ is not a DAG[3], and the iteration may not converge in a finite number of steps. In order to avoid such situation, as well as an undesirable retro-feeding effect of the expansion (e.g. the initial non-zero user preferences should not be increased by the expansion), $\hat{T}$ is made acyclic before starting the iteration. This is achieved by removing the appropriate arcs before expansion, namely the ones that eliminate all cycles and at the same time maximize the resulting $E(X)$. The fact that the above equation converges is assured by this transformation, applied before the contextualization step and described in the following algorithm.

In the above formulas, the '$\circ$' sign denotes the fuzzy composition between a fuzzy relation and a fuzzy set. In the general case and given a fuzzy relation $\mathcal{R} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ and a fuzzy set $\mathcal{A}' : \mathcal{X} \rightarrow [0, 1]$, the fuzzy composition is defined as:

$$\mathcal{B}' = \mathcal{A}' \circ \mathcal{R} : \mathcal{Y} \rightarrow [0, 1] \tag{39}$$

and:

$$\mathcal{B}' = \bigcup_{x \in \mathcal{X}} (\mathcal{A}' \cap \mathcal{R}) \quad or \quad \mu_{\mathcal{B}'}(y) = u_{x \in \mathcal{X}}(t(\mu_{\mathcal{A}'}(x), \mu_{\mathcal{R}}(x, y))) \tag{40}$$

where t and u are a fuzzy t-norm and a fuzzy t-conorm, respectively.

The expansion operation described above is implemented in our system by the following procedure:

```
expand set (X; E(X))
  for x ∈ S do
    E(X)(x)←X(x)
    in_path[x]←false
  for x ∈ supp(X) do
    expand_concept(x, 0)
expand concept(x, prev_x)
  in_path[x]←true
  for y ∈ {z ∈ S|R̂(x, z) > 0} do
    if not in_path[y] and X(y) = 0 and E(X)(y) < 1 then
```

[3] Directed Acyclic Graph

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

$prev_y \leftarrow E(X)(y)$
$E(X)(y) \leftarrow (E(X)(y) - \hat{R}(x,y) * prev_x / (1 - \hat{R}(x,y) * prev_x)$ /* Undo last update from x*/
$E(X)(y) \leftarrow E(x)(y) + (1 - E(X)(y)) * \hat{R}(x,y) * E(X)(x)$
if $E(X)(y) > \varepsilon$ then expand concept(y, $prev_y$)
in_path[x] $\leftarrow$ *false*

The algorithm is shown here in pseudocode and in a recursive version for the sake of readability, but it has been implemented in practice as an iteration, using a stack. The set $supp(X)$ denotes the crisp support of X, that is, the set $\{x \in S | X(x) > 0\}$. The in_path[x] attribute in the expand_concept procedure is what makes the proper $\hat{R}$ arcs be removed, that is, the arcs by which a concept x would contribute to its own expansion are temporarily deactivated in the iteration. The ε value is a minimum threshold below which the value of $E(X)(x)$ is not expanded to the semantic neighborhood of x.

It can be shown that the above algorithm achieves the expansion method with $O(|supp(X)| \cdot |S| \cdot |supp(\hat{R})|)$ complexity, where $supp(\hat{R})$ denotes the crisp support of $\hat{R}$, that is, the set of pairs $(x,y) \in S \times S$ with $\hat{R}(x,y) > 0$. However, in practice the cost is much lower, since $E(X)(y)$ quickly decays below ε as y gets farther away from the initial concepts having $X(x) > 0$ (where 'far' refers to the number of $\hat{R}$ arcs needed to reach y from this set). Our experiments, reported in Section 7 of this manuscript, show that the time spent in the expansion itself is irrelevant compared to the cost of other operations of the program, such as accessing the KB.

The extended runtime context $E(\hat{C}_t)$, as well as the extended set of user preferences $E(P)$, are computed based on equation (38) and consequently the precise expression of the contextualized user preferences CP_t is given by the algebraic product of the two fuzzy sets:

$$CP_t = E(P)E(\hat{C}_t) \tag{41}$$

Now CP_t can be interpreted as a combined measure of the likelihood that a concept is preferred and how relevant the concept is to the current context. Note that this fuzzy set is in fact dependent on both user and time, that is, $CP_t(\hat{u})$, and that at this point we have achieved a contextual-preference mapping as defined in Section 3.1, namely $\Phi(P, \hat{C}_t) = CP_t$, where $P \vDash \Phi(P, \hat{C}_t))$, since $CP_t > P$ only when $E(P)$ has been derived from P and $CP_t < E(P)$.

5.2 *Ranking*

Finally, given a document $d \in \mathcal{D}$ ($\mathcal{D}$ being the set of all documents in the retrieval space, as already introduced in Subsection 2), the predicted interest (to which we shall refer as personal ranking measure, $r_P(d,t)$ of the user $\hat{u}$ for d at a given instant t in a session is measured as a value in the interval [0,1], based on his/her preferences P and computed by:

$$r_P(d,t) = cos(S_d, CP_{t-1}) \tag{42}$$

where S_d is the fuzzy set of concepts associated to $d \in \mathcal{D}$ and CP_{t-1} the set of contextualized preferences obtained from the previous Subsection 5.1. Equation (42) holds as, according to (Egghe & Michel, 2003), given two fuzzy sets $\mathcal{X}, \mathcal{Y} \in \mathcal{F}_S$, their cosine similarity measure is defined as:

$$cos(\mathcal{X}, \mathcal{Y}) = \frac{|\mathcal{X} \cap \mathcal{Y}|}{\sqrt{|\mathcal{X}||\mathcal{Y}|}} \tag{43}$$

To interpret equation (43) we provide the extension of the cardinality of a crisp set to the fuzzy case, defined as follows:

$$|\mathcal{X}| = \sum_{x \in \mathcal{X}} P_{\mathcal{X}}(x) \tag{44}$$

and utilize *min* in the fuzzy intersection of the fuzzy sets $\mathcal{X}$ and $\mathcal{Y}$. Thus:

$$r_P(d,t) = \frac{|S_d \cap CP_{t-1}|}{\sqrt{|S_d||CP_{t-1}|}} \tag{45}$$

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

In the context of a content-retrieval system, where users retrieve contents by issuing explicit requests and queries, the $r_P(d, t)$ measure is combined with query-dependent, user-neutral search result rank values, to produce the final, contextually personalized, rank score for the document.

The final, contextually personalized, rank score $r(d, t)$ for the document d is then given by:

$$r(d, t) = f(r_P(d, t), r_S(d, t)) \tag{46}$$

The similarity measure $r_S(d, t)$ stands for any ranking technique to rank a document d with respect to a query or request at a given time t. $r_S(d, t)$ is computed according to the given possible user queries described in Section 4.2. For instance, in the case of keyword-based queries we have:

$$r_S(d, t) = cos(S_d, Q_t^k) \tag{47}$$

or in the case of topic browsing, the degree to which document d is classified to topic Q_t^b is given by:

$$r_S(d, t) = G_d(z) \tag{48}$$

where $z = Q_t^b$ is the specific topic and $G_d = G(S_d)$ is the topic-classification output of the topic-classification process. The latter follows the same guidelines as analytically described in Section 4.4, that is, both processes of user-preferences extraction and topic classification implement the same algorithm and can be defined in terms of the same function G:

- $G(S_d) = G_d$ provides the fuzzy set of all topics associated to the specific document d,
- z corresponds to a specific topic element of this fuzzy set, and
- $G_d(z)$ denotes the degree to which z belongs to G_d or in other words its membership degree.

Documents are ranked according to their similarity to the predefined topic of search, whereas in the case of a single-view document query, the requested document d is simply presented to the user.

In general, the score (46) can be used to introduce a personalized bias into any ranking technique that computes $r_S(d, t)$, which could be image-based, ontology-based, relevance-feedback based, etc. The combination function f can be defined for instance as a linear combination $f(x, y) = \lambda \cdot \bar{x} + (1-\lambda) \cdot \bar{y}$. The term λ is the personalization factor that shall determine the degree of personalization applied to the search result ranking, ranging from $\lambda = 0$ producing no personalization at all, to $\lambda = 1$, where the query is ignored and results are ranked only on the basis of global user interests. As a general rule, λ should decrease with the degree of uncertainty about user preferences, and increase with the degree of uncertainty in the query. The problem of how to set the value dynamically is addressed by the authors in the study of Gauch *et al.* (2004), where the reader is encouraged to find further details. $\bar{x}$ and $\bar{y}$ denote the normalization of the score values x and y, which is needed before the combination to ensure that they range on the same scale. The final value $r(d, t)$ determines the position of each document d in the final ranking in the personalized search result presented to the user.

6 A use case

As an illustration of the application of the contextual-personalization techniques, consider the following scenario: Elli is subscribed to an economic news content provider. She works for a major food company, so she has preferences for news related to companies of this sector, but she also tries to be up-to-date in the technological domain, as her company, as well as her personal interests, are trying to apply the latest technologies in order to optimize the food production chain and be technologically up-to-date in the computer world.

Elli is planning a trip to Tokyo and Kyoto, in Japan. Her goal is to take ideas from different production chains of several Japanese partner companies. She has to document about different companies in Japan, so she accesses the content provider and begins a search session. The scenario

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Table 2 Concept names and mnemonics

Concept	Mnemonic	Concept	Mnemonic
Food companies	**fcp**	Japan Tobacco Inc.	jti
Mc' Donalds	mcd	Big Mac	bgm
Yamazaki Baking Co.	yam	**Technology companies**	**tcp**
Microsoft Corp.	**msc**	**Apple Computers Inc.**	**apl**
Microsoft Office Suite	ofc	Personal Computer	pcm
Macintosh	mac	**Linux Community**	**lnx**
Tux	tux	X Windows System	xws
Programming shell	shl	Windows media player	wmp
Microsoft Visio	vso	Windows Mathtype	mtp
Dunkin Donuts	dnk	Coca Cola	cok
Food, Beverage & Tobacco Sector	fbt	Makoto Tajima	mkt
Microsoft	mis	Apple	ape
McDonald's Corp.	mdc	Macintosh G3	mcg

Note: Topics are shown in boldface.

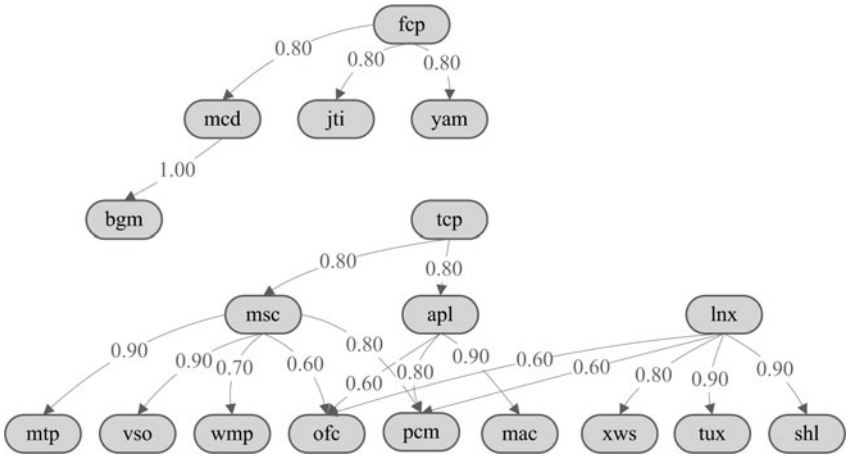


Figure 4 Example of T relation construction

comprises the documents, a set of concepts S and relations T and $\hat{T}$ defined on S, as described in the previous subsections. The first step of our methodology is the estimation of the set of off-line user preferences P for Elli, that is, the estimation of the weighted semantic interests for domain concepts of the ontology included in Elli's profile. The set of concepts together with their mnemonics is presented in Table 2, relation T in Figure 4, and its values in Table 3. The set of concepts includes several companies from the food, beverage, and tobacco sector and also several technological companies. Only the relevant concepts have been included, together with their degrees of membership. This would lead to the definition of the P fuzzy set for Elli, as described in the following section.

Relation elements that are implied by transitivity are omitted for the sake of clarity; sup-product is assumed for transitivity and the t-norm used for the transitive closure of relation T is Yager's t-norm[4] with parameter $p = 3$. In addition, the co-norm used in equation (16) is the bounded sum, while in (30), the t-norm used is the product and the standard co-norm *max* is utilized for final preference extraction. Finally, the threshold used for the termination criterion of the clustering algorithm is 0.3.

[4] $T_p^{\mathrm{Y}}(x,y) = \max\left(0, 1-((1-x)^p+(1-y)^p)^{1/p}\right), \quad for \quad 0<p<+\infty$

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Table 3 Part of the taxonomic relation T

s_1	s_2	T	s_1	s_2	T	s_1	s_2	T
msc	wmp	0.70	tcp	apl	0.80	tcp	msc	0.80
fcp	mcd	0.80	lnx	xws	0.80	mcd	bgm	1.00
fcp	jti	0.80	apl	mac	0.90	lnx	ofc	0.60
fcp	yam	0.80	apl	pcm	0.80	apl	ofc	0.60
msc	mtp	0.90	lnx	pcm	0.60	lnx	tux	0.90
msc	vso	0.90	msc	ofc	0.60	lnx	shl	0.90
msc	pcm	0.80						

The semantic indexing is represented as:

$$I(s_j, d) = \text{pcm}/0.9 + \text{dnk}/0.8 + \text{ofc}/0.9 + \text{mac}/1 + \text{jti}/0.4 \tag{49}$$

The concept-clustering process results into 3 crisp clusters:

$$K' = \{k_1', k_2', k_3'\} = \{(pcm, mac, ofc), dnk, jti\} \tag{50}$$

Due to the simplicity of the content in this first session of Elli and the small amount of its detected concepts, the use of the context-based classifier introduced in Subsection 4.4.3 does not lead to an expansion of the detected crisp clusters, that is, to include other concepts. This is expected by observing the structure of the T relation in Figure 4, since the semantics of all concepts imply either a full or a absolutely absent relation. We further adjust the degrees of membership for these clusters, using the product t-norm and according to equation (30), as follows:

$$k_1 = pcm/0.9 + mac/1.0 + ofc/0.9 \tag{51}$$

$$k_2 = dnk/1.0 \tag{52}$$

$$k_3 = jti/1.0 \tag{53}$$

Each one of the above clusters corresponds to one of the distinct user preferences associated to Elli and in order to determine them we have considered both the scalar cardinality of each cluster, as well as its context. More specifically, for each cluster we have:

$$h(k_1) = 1.0 \quad and \quad |k_1| = 3 \tag{54}$$

$$h(k_2) = 0.8 \quad and \quad |k_2| = 1 \tag{55}$$

$$h(k_3) = 0.4 \quad and \quad |k_3| = 1 \tag{56}$$

Their context is calculated as:

$$C(k_1) = apl/0.6 + tcp/0.58 \tag{57}$$

$$C(k_2) = \varnothing \tag{58}$$

$$C(k_3) = fcp/0.8 \tag{59}$$

Applying the weak modifier $w(a) = \sqrt{a}$, we obtain:

$$w(C(k_1)) = \sqrt{C(k_1)} = apl/0.77 + tcp/0.76 \tag{60}$$

$$w(C(k_2)) = \varnothing \tag{61}$$

$$w(C(k_3)) = \sqrt{C(k_3)} = fcp/0.89 \tag{62}$$

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

As described in Subsection 4.4.4 clusters k_2 and k_3 are of extremely low cardinality and thus contain misleading concepts. After adjusting the membership degrees of the clusters according to their scalar cardinalities using the triangular fuzzy number$(1.3, 3, \infty)$, both clusters are ignored in the estimation of $\{W(k_i)\}$. Finally, the set of user preferences associated to Elli at this point is given by:

$$\{W(k_i)\} = \underset{k \in K}{\cup} W(k) = W(k_1) = apl/0.77 + tcp/0.76 \tag{63}$$

As the time goes by and without any loss of generality, we may assume that the proposed framework has learned some of Elli's preferences over time following the previous methodology and her related queries, that is, Elli's profile is now enhanced and includes the weighted semantic interests for domain concepts of the ontology shown next. These include several companies from the food, beverage, and tobacco sector and also several technological companies. Again, only the relevant concepts have been included together with their degrees of membership. This would lead us to consider the P fuzzy set, as defined in Section 4.1, as:

$$P = \{yam/0.85 + jti/0.92 + mcd/0.74 + apl/0.77 + msc/0.66 + tcp/0.76\} \tag{64}$$

In order to proceed to the next step of identifying the runtime context within our approach, we utilize the relationships between the concepts of P, as they are defined according to the relation $\hat{T}$ and are exemplified in Figure 5. $\hat{T}$-relation values were initially set by exploiting all available information in the KB and by manually analyzing and checking the effect of propagation on a list of use cases for the combined relation, and was tuned empirically afterwards. Investigating methods for automatically learning of values is an open research direction for our future work.

When Elli enters a query (the query-based search engine can be seen essentially as a black box for our technique), the personalization system adapts the result ranking to Elli's preferences by combining the query-based similarity measure $r_S(d, t)$ and the preference-based $r_P(d, t)$ scores for each document d that matches the query, as described in Subsection 5.2. At this point, the adaptation is not contextualized, since Elli has just started the search session, and the runtime context is still empty (i.e. at $t = 0$, $\widehat{C}_0 = \emptyset$).

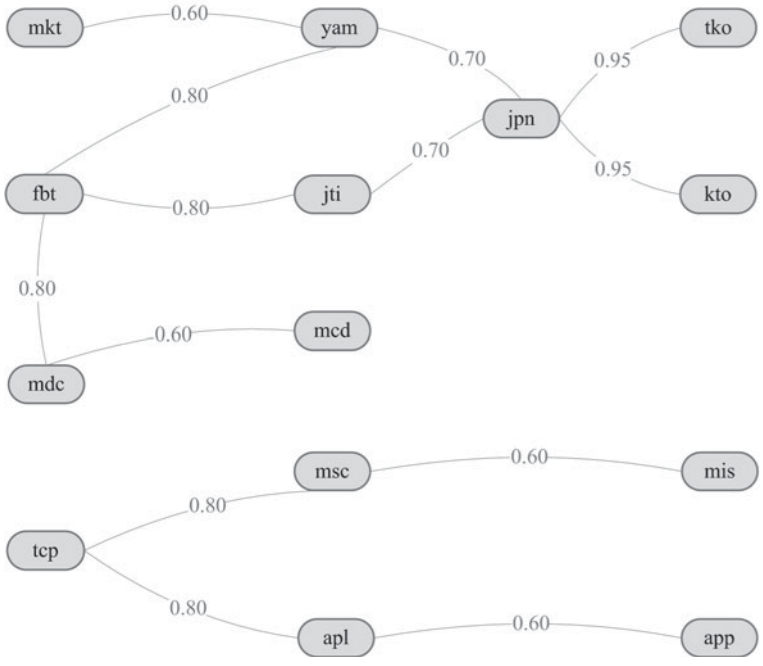


Figure 5 A subset of the $\hat{T}$ relation connecting the concepts involved in the expansion of Elli's runtime context

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Table 4 Part of relation $\hat{T}$

s_1	s_2	$\hat{T}$	s_1	s_2	$\hat{T}$	s_1	s_2	$\hat{T}$
tko	jpn	0.95	kto	jpn	0.95	jpn	yam	0.70
jpn	jti	0.70	jti	fbt	0.80	fbt	yam	0.80
mkt	yam	0.60	fbt	mdc	0.80	mdc	mcd	0.60
tcp	msc	0.80	msc	mis	0.60	tcp	apl	0.80
apl	app	0.60						

But now suppose that the need of information expressed in the first query is somehow related to the concepts Tokyo (*tko*) and Kyoto (*kto*), as Elli wants to find information about the cities she's visiting. Thus, she opens and saves some general information documents about the living and economic style of these two cities. As a result, the system builds a runtime context out of the metadata of the selected documents and the executed query, which forms the $\hat{C}$ fuzzy set:

$$\hat{C}_1 = \{tko/1.0 + kto/1.0\} \tag{65}$$

Now, Elli wants to see some general information about Japanese companies. The contextualization mechanism comes into place, as follows.

1. First, the context set is expanded through semantic relations from the initial context, adding more weighted concepts, shown in bold in the $E(\hat{C}_t)$ fuzzy set for Elli, following the notation used in Subsection 5.1 and the part of relation $\hat{T}$ illustrated in Table 4.
   By applying the semantic-extension methodology described in Subsection 5.1 and by using in this case the algebraic sum and the algebraic product as the fuzzy *t*-conorm and fuzzy *t*-norm in equation (40), respectively, we obtain:

$$\begin{aligned} E(\hat{C}_1) = \{&tko/1.00 + kto/1.00 + \boldsymbol{jpn/1.00} + \boldsymbol{jti/0.89} + \boldsymbol{yam/0.64} \\ &+ \boldsymbol{mkt/0.64} + \boldsymbol{fbt/0.78} + \boldsymbol{mdc/0.67} + \boldsymbol{mcd/0.45}\} \end{aligned} \tag{66}$$

2. Similarly to the above process, Elli's initial preferences are extended through semantic relations from her initial ones. The expanded preferences stored in the $E(P)$ fuzzy set are the following, where the new/updated concepts are in bold:

$$\begin{aligned} E(P) = \{&yam/0.85 + jti/0.92 + tcp/0.76 + msc/0.66 + mcd/0.74 + apl/0.77 \\ &+ \boldsymbol{jpnl/0.89} + \boldsymbol{tko/0.86} + \boldsymbol{kto/0.86} + \boldsymbol{fbt/0.95} + \boldsymbol{mkt/0.83} + \boldsymbol{mdc/0.90} \\ &+ \boldsymbol{mis/0.73} + \boldsymbol{app/0.75}\} \end{aligned} \tag{67}$$

3. The contextualized preferences are computed as described in Subsection 5.1, yielding the following *CP* fuzzy set (concepts with membership degree equal to 0 are omitted):

$$\begin{aligned} CP_1 = \{&yam/0.54 + jti/0.82 + mcd/0.33 + jpn/0.89 + tko/0.86 \\ &+ kto/0.86 + fbt/0.74 + mkt/0.53 + mdc/0.60\} \end{aligned} \tag{68}$$

   Comparing this to the initial preferences in Elli's profile, we can see that Microsoft Corp., Apple Computers Inc., and Technology companies are disregarded as out-of-context preferences, whereas Japan Tobacco Inc., McDonald's, and Yamazaki Baking Co. have been retained because they are semantically related both to the initial Elli's preferences (food sector), and to the current context (Japan). Moreover, Japan, Tokyo, and Kyoto have been added as instructed by the initial context. It is worth noting that Japan was not included in the initial runtime context and is added because of relation $\hat{T}$. Finally, Makoto Tajima, McDonald's Corp., and Food, Beverage & Tobacco Sector are also included with lower degrees of membership as in-context user preferences.
4. Using the contextualized preferences above, a different personalized ranking is computed in response to the current user query based on the $E(\hat{C}_1)$ fuzzy set, instead of the basic *P* preference fuzzy set.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

7 Experimental results

The contextualization techniques described in the previous sections have been implemented in an experimental prototype, and tested on a medium-scale corpus. Evaluating personalization is known to be a difficult and expensive task (Wilkinson & Wu, 2004; Rajagopalan & Deshmukh, 2005). In order to measure how much better a retrieval system can perform with the proposed techniques than without them, it is necessary to compare the performance of retrieval (1) without personalization, (2) with simple personalization, and (3) with contextual personalization. The standard evaluation measures from the IR field require the availability of manual content ratings with respect to (1) query relevance, (2) query relevance and general user preference (i.e. regardless of the task at hand), and (3) query relevance and specific user preference (i.e. constrained to the context of his/her task).

For this purpose, we have conducted two sets of experiments. Both are based on the same search space corpus, consisting of 145316 documents (445MB) from the CNN Web site (http://dmoz.org/News/Online_Archives/CNN.com), plus the KIM domain ontology and KB (Kiryakov *et al.*, 2004), publicly available as part of the KIM Platform, developed by Ontotext Lab, with minor extensions. The KB contains a total of 281 RDF (Klyne *et al.*, 2004) classes, 138 properties, 35689 instances, and 465848 sentences. The CNN documents are annotated with KB concepts, amounting to over three million annotations in total. The fuzzy relation values were first set manually on an intuitive basis, and tuned empirically afterwards by running a few trials. The user-neutral retrieval system used for this experiment is a semantic search engine developed by the authors (Castells *et al.*, 2007). This engine has been shown to have better performance than a traditional keyword-based system such as the Jakarta Lucene library (http://lucene.apache.org), when ontological knowledge is available (see Castells *et al.*, 2007), thus providing a harder baseline for our evaluation. The experiments reported here test only the performance of the retrieval phase, taking predefined user preferences as a starting point. User preferences are simulated in the first set of experiments, and manually provided by real users in the second set.

Since the contextualization techniques are applied in the course of a session, one way to evaluate them is to define a sequence of steps where the techniques are put to work. This is the approach followed in the first set of experiments, for which we have built a testbed consisting of a fixed set of hypothetic context situations, detailed step by step. The testbed comprises ten short use cases, including the one explained in the previous section. Each scenario consists of a sequence of user actions defined a priori, including queries and clicks on search results. When it comes to compute precision and recall measures, this approach makes it difficult to get detailed user assessments (ground truth), because of the effort and difficulty involved in assessing results under a large set of artificial, complex, and demanding assumptions, imposed to the human judges. Therefore, we have rated the document/query/preference/context tuples manually, based on hypothetic users, for whom user profiles are simulated. Although subjective, this approach allows meaningful observations, and testing the feasibility, soundness, and technical validity of the defined models and algorithms. These results are complemented with a more objective, though less detailed evaluation with real users which will be described after this.

Figure 6a shows the results of this experimental approach for the use case described in the previous section. This is a clear example where personalization alone would not give better results, or would even perform worse than non-adaptive retrieval (see the drop of precision for recall between 0.1 and 0.4 in Figure 6a), because irrelevant long-term preferences (such as, in the example, technological companies which are not related to the current user focus on Japan-based companies) would get in the way of the user. The experiment shows how our contextualization approach can avoid this effect and significantly enhance personalization by removing such out-of-context user interests and leaving the ones that are indeed relevant in the ongoing course of action.

It can also be observed that the contextualization technique consistently results in better performance with respect to simple personalization, as can be seen in Figure 6b, which shows the average results over ten use cases, and Figure 7, depicting the average precision histogram comparing the contextualized vs. non-contextualized personalization at retrieval time.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

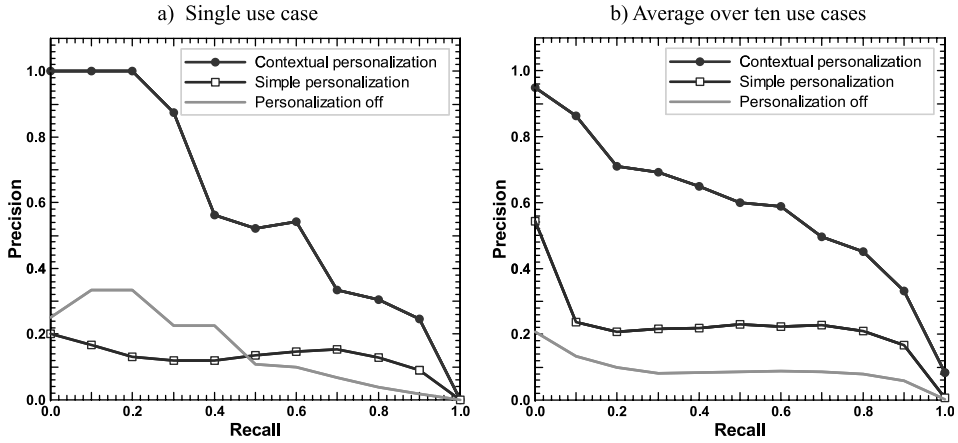


Figure 6 Comparative performance of personalized search with and without contextualization, showing the precision vs. recall curve for (a) one of the scenarios, and (b) the average over 10 scenarios. The results in graphic (a) correspond to the query 'Companies based in any Japanese region', from the use case described in Section 6

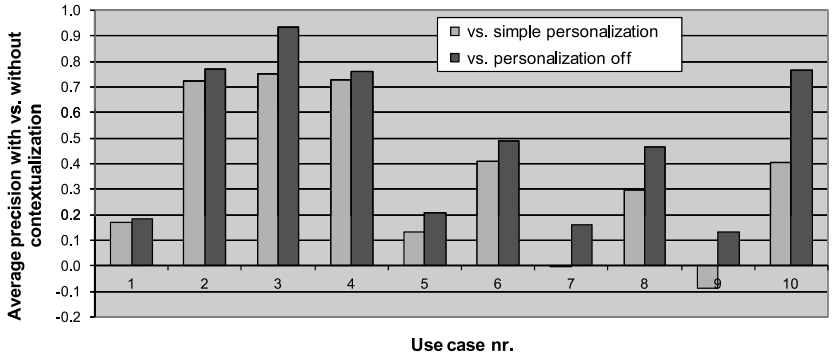


Figure 7 Comparative precision histogram of personalized search with and without contextualization for the ten use cases. The light-colored bars compare personalized retrieval in context vs. simple personalized retrieval without context, and the dark-colored ones compare personalized retrieval in context vs. retrieval without personalization

In the second approach, real human subjects are given three different retrieval tasks, each expressing a specific information need, so that users are given the goal of finding as many documents as possible which fulfill the given needs. In this experiment, the sequence of actions is not fixed as in the previous one, but is defined with full freedom by users as they seek to achieve the proposed tasks. The semantic-query capabilities are disabled this time, to avoid complexities in the interaction with users which could distort the results. Users enter their searches as plain keyword-based queries, and the Lucene library is used as the primary search engine (providing the user-neutral $r_S(d, t)$ values described in Subsection 5.2).

A total of 18 subjects were selected for the experiment, all of them being PhD students from the authors' institutions. Three tasks were set up for the experiment, which can be briefly summarized as:[5]

1. News about agreements between companies.
2. Presentations of new electronic products.
3. Information about cities hosting a motor sports event.

[5] In practice the users are given a more detailed and verbose description of the topic, in order to define it as precisely as possible and to avoid ambiguities.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Each task was tested (a) with contextual personalization, (b) with simple personalization, and (c) without personalization. In order for users not to repeat the same task twice or more, each of the three modes was used with six users (3 modes × 6 users = 18 tests for each task), in such a way that each user tried each of the three modes (a), (b), and (c), exactly once, following a Latin square experimental design. This way, each mode is tried exactly 18 times: once for each user, and 6 times for each task, in such a way that neither mode is harmed or favored by different task difficulty or user skills. User preferences are obtained manually from users by asking them to explicitly rate a predefined list of domain concepts at the beginning of the session.

The relevant documents for each task are marked beforehand by an expert (a role that we played ourselves), so that users are relieved from providing extensive relevance judgments. However, users are encouraged to open the documents that seem more relevant according to their subjective interests, in order to provide the system with more contextual tips. Context information is gathered based on concepts annotating such selected results, and the concepts that are related to the keywords in user queries (using the keyword-concept mapping provided in the KIM KB).

At the end of every task the systems ask the user to mark the documents in the final result set as related or unrelated to her particular interests *and* the search task. For the computation of precision and recall after the experiment logs were collected, the following two simplifications are made for each interactive sequence (i.e. for each task and user):

- The search space is simplified to be the set of all documents that have been returned by the system at some point in the iterative-retrieval process for the task conducted by this user.
- The set of relevant documents is taken to be the intersection of the documents in the search space marked as relevant for the task by the expert judgment, and the ones marked by the user according to her particular interests.

Figure 8 shows the results obtained with this setup and methodology. The curve on the left of this figure shows a clear improvement at high precision levels by the contextualization technique both with respect to simple personalization and no personalization, an improvement which decreases at higher recall levels. The improvement by the contextual personalization is similarly apparent in the cut-off precision curve, especially in the top 10 results. Personalization alone achieves considerably lower precision on the top documents, showing that the contextualization technique avoids an important number of false positives which may occur when user preferences are considered out of context. The mean average precision values shown in Table 5 for contextual, simple, and no personalization in this experiment confirm that our technique globally performs clearly above the two baselines.

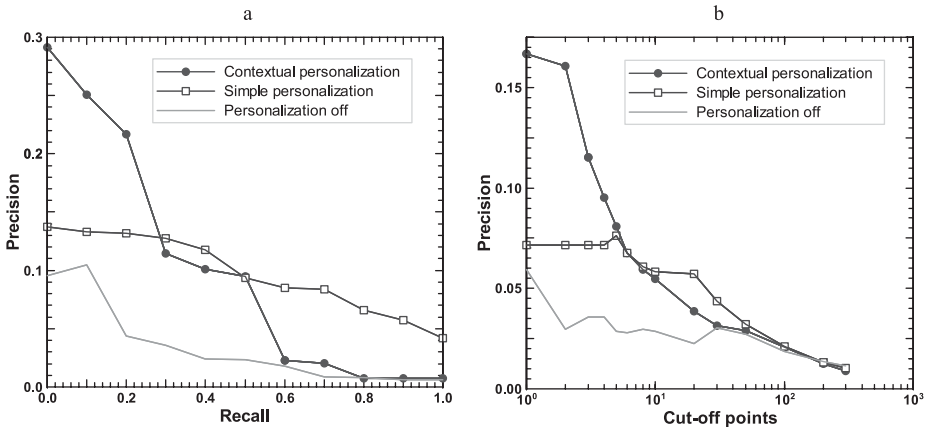


Figure 8 Comparative performance of personalized search with and without contextualization tested with 18 subjects on three proposed tasks. The graphics show (a) the precision vs. recall curve, and (b) the precision at cut-off points. The results are averaged over the set of all users and tasks

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Table 5 Mean average precision for each of the three retrieval modes

Retrieval mode	MAP
Contextual personalization	0.1353
Simple personalization	0.1061
Personalization off	0.0463

Most cases where our technique performed worse were due to a lack of information in the KB, as a result of which the system did not find that certain user preferences were indeed related to the context. Another limitation of our approach is that it assumes that consecutive user queries tend to be related, which does not hold when sudden changes of user focus occur. However, not only the general improvements pay off on average, but the potential performance decay in such cases disappears after two or three queries, since the weight of contextual concepts decreases exponentially as the user keeps interacting with the system, as explained in Subsection 5.1. Nonetheless, as future work, it would be possible to enhance our approach by assessing the semantic distance between user requests, and clustering the context into cohesive subsets, leading to an even finer contextualization.

8 Conclusions

Context is an increasingly common notion in IR, and has been identified as a major challenge in the field (Allan *et al.*, 2002). This is not surprising since it has long been acknowledged that the whole notion of relevance, at the core of IR, is strongly dependent on context—in fact it can hardly make sense out of it. Several authors in the IR field have explored approaches that are similar to ours in that they find indirect evidence of searcher interests by extracting implicit meanings in information objects manipulated by users in their retrieval tasks. A key differentiating aspect in our approach is the use of semantic concepts, rather than terms (i.e. strings), for the representation of these contextual meanings, and the exploitation of explicit ontology-based information attached to the concepts, available in a knowledge base.

Ontologies provide indeed a powerful vehicle to represent a wide range of descriptions of content qualities and user wants, in a way allowing to relate what the user likes to what he is currently asking for and what he is paying attention to, and match this to what a content provides, in a fairly precise way. The formal information (such as concept classification and explicit semantic relations) provided in full-fledged domain ontologies enables more accurate and reliable results than the statistical techniques used in previous proposals, which for example, estimate term similarities out of their statistic co-occurrence in a content corpus. Complementing the ontology-based approach with fuzzy representations of user interests, user context, and content meaning, it brings to bear additional capabilities from available fuzzy theory and models, to tackle the imprecision and uncertainty involved in the meanings and phenomena under study.

Acknowledgments

This research was partially supported by the European Commission under contracts FP6-001765 aceMedia and FP6-027685 MESH. The expressed content is the view of the authors but not necessarily the view of the aceMedia or MESH projects as a whole.

References

Al-Khatib, W., Day, Y. F., Ghafoor, A. and Berra, P. B. 1999 Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering* **11**(1), 64–80.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Allan, J. *et al*. 2002 *Challenges in information retrieval and language modelling, Report of Workshop held at the University of Massachusetts, Amherst*. *SIGIR Forum* **37**(1), 31–47.

Benkhalifa, M., Bensaid, A. and Mouradi, A. 1999 Text categorization using the semi-supervised fuzzy c-means algorithm. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS 1999)*, New York, USA, pp. 561–565.

Bharat, K. 2000 SearchPad: Explicit capture of search context to support web search. In *Proceedings of the 9th International World Wide Web Conference (WWW9)*, Amsterdam, The Netherlands, pp. 493–501.

Brown, P. J., Bovey, J. and Chen, X. 1997 Context-aware applications: from the laboratory to the marketplace. *IEEE Personal Communications* **4**(5), 58–64.

Campbell, I. and Van Rijsbergen, C. J. 1996 The ostensive model of developing information needs. In *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science (CoLIS 1996)*, Copenhagen, Denmark, pp. 251–268.

Castells, P., Fernandez, M. and Vallet, D. 2007 An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering* **19**(2), 261–272.

Castells, P., Fernandez, M., Vallet, D., Mylonas, Ph. and Avrithis, Y. 2005 Self-tuning personalized information retrieval in an ontology-based framework. In *Proceedings of the 1st International Workshop on Web Semantics*, Agia Napa, Cyprus, Springer Verlag LNCS vol. 3762, pp. 977–986.

Chang, C. H. and Hsu, C. C. 1998 Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval. *Computer Networks and ISDN Systems* **30**(1–7), 621–623.

Coutaz, J., Crowley, J., Dobson, S. and Garlan, D. Context is key. *Communications of the ACM* **48**(3), 49–53.

Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V. K. and Strintzis, M. G. 2005 Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **15**(10), 1210–1224.

Dill, S. *et al*. 2003 A case for automated large scale semantic annotation. *Journal of Web Semantics* **1**(1), 115–132.

Edmonds, B. 1999 The pragmatic roots of context. In *Proceedings of the 2nd International and Interdisciplinary Conference on Modeling and Using Context*, Trento, Italy, Springer Verlag LNAI vol. 1688, pp. 119–132.

Egghe, L. and Michel, C. 2003 *Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques*. *Information Processing and Management* **39**(5), 771–807.

Ehrig, M. and Sure, Y. 2004 Ontology mapping—an integrated approach. In *Proceedings of the 1st European Semantic Web Symposium (ESWS 2004)*, Heraklion, Greece, Springer Verlag LNCS vol. 3053, pp. 76–91.

Euzenat, J. 2004 Evaluating ontology alignment methods. In *Proceedings of the Dagstuhl Seminar on Semantic Interoperability and Integration*, Wadern, Germany, pp. 47–50.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2002 Placing search in context: The concept revisited. *ACM Transactions on Information Systems* **20**(1), 116–131.

Gauch, S., Chaffee, J. and Pretschner, A. 2004 Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems* **1**(3–4), 219–234.

Haveliwala, T. H. Topic-sensitive pagerank. In *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, USA, pp. 517–526.

Heer, J., Newberger, A., Beckmann, C. and Hong, J. 2003 Liquid: context-aware distributed queries. In *Proceedings of the 5th International Conference on Ubiquitous Computing (UbiComp 2003)*, Seattle, Washington, USA, pp. 140–148.

Hong, J. I. and Landay, J. A. 2001 An infrastructure approach to context-aware computing. *Human-Computer Interaction* **16**(2–4), 87–96.

ISO/IEC FDIS 15938-5, ISO/IEC JTC 1/SC 29 M 4242. 2001 *Information Technology—Multimedia Content Description Interface—Part 5: Multimedia Description Schemes*, pp. 442–448.

Jeh, G. and Widom, J. 2003 Scaling personalized web search. In *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*, Budapest, Hungary, pp. 271–279.

Kalfoglou, Y. and Schorlemmer, M. 2003 Ontology mapping: the state of the art. *Knowledge Engineering Review* **18**(1), 1–31.

Kelly, D. and Teevan, J. Implicit feedback for inferring user preference. *SIGIR Forum* **32**(2), 18–28.

Kim, H. and Chan, P. 2003 Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2003)*, Miami, Florida, USA, pp. 101–108.

Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. 2004 Semantic annotation, indexing, and retrieval. *Journal of Web Sematics* **2**(1), 47–49.

Klir, G. and Bo, Y. 1995 *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. New Jersey, USA: Prentice Hall.

Klyne, G., Carrol, J. J., and McBride, B. 2004 *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press

Kohavi, R. and Sommerfield, D. 1995 Feature subset selection using the wrapper model: overfitting and dynamic search space topology. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1995)*, Montréal, Canada.

Kraft, D. H., Bordogna, G., and Passi, G. 1998 Information retrieval systems: where is the fuzz?, In *Proceedings of IEEE International Conference on Fuzzy Systems*, Anchorage, Alaska, USA, pp. 1367–1372.

Lawrence, S. 2000 Context in web search. *IEEE Data Engineering Bulletin* **23**(3), 25–32.

Lewis, D. 1980 Index, context, and content. In Kanger, S. and Ohman, S. (Eds.), *Philosophy and Grammar*, pp. 79–100, D. Reidel Publishing Company, Dordrecht, Holland.

Liu, F., Yu, C. and Meng, W. 2004 Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering* **16**(1), 28–40.

McCarthy, J. 1993 Notes on formalizing context. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993)*, Chambéry, France, pp. 81–98.

Micarelli, A. and Sciarrone, F. 2004 Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Modelling and User-Adapted Interaction* **14**(2–3), 159–200.

Miyamoto, S. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht Boston London: Kluwer Academic Publishers.

Mylonas, Ph. and Avrithis, Y. Context modeling for multimedia analysis and use. In *Proceedings of the 5th International and Interdisciplinary Conference on Modeling and Using Context (Context 2005)*, Paris, France.

Mylonas, Ph., Wallace, M., and Kollias, S. Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering. In Vouros, G. A. and Panayiotopoulos, T. (Eds.), *Methods and Applications of Artificial Intelligence*, Springer Verlag LNCS vol. 3025, pp. 191–200.

Noy, N., Semantic integration: a survey of ontology-based approaches, *Sigmod Record* **33**(4), 65–70.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D. and Kirilov, A. 2004 KIM—A semantic platform for information extraction and retrieval. *Journal of Natural Language Engineering* **10**(3–4), 375–392.

Rajagopalan, B. and Deshmukh, A. Evaluation of online personalization systems: a survey of evaluation schemes and a knowledge-based approach. *Journal of Electronic Commerce Research* **6**(2), 112–122.

Rocchio, J. 1971 Relevance feedback information retrieval. In Salton, G. (Eds.), *The Smart Retrieval System-Experiments in Automatic Document Processing*, Kansas City, Missouri, USA: Prentice Hall, pp. 313–323.

Salton, G. and McGill, M. 1983 *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Shen, X., Tan, B., and Zhai, C. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, Salvador, Brazil, pp. 43–50.

Staab, S. and Studer, R. (Eds.) 2004 *Handbook on Ontologies*. Heidelberg, Berlin and New York: Springer Verlag.

Theodoridis, S. and Koutroumbas, K. 1998 *Pattern Recognition*. Academic Press, San Diego, CA, USA.

Vallet, D., Mylonas, Ph., Corella, M. A., Fuentes, J. M., Castells, P. and Avrithis, Y. 2005 A semantically-enhanced personalization framework for knowledge-driven media services. In *Proceedings of IADIS International Conference on WWW / Internet (ICWI 2005)*, Lisbon, Portugal.

van Eijck, J. 2000 On the proper treatment of context in NL. In Monachesi, P. (Ed.), Computational Linguistics in the Netherlands 1999, *Selected Papers form the 10th CLIN Meeting*, Utrecht, The Netherlands.

White, R. W., Jose, J. M., van Rijsbergen, C. J. and Ruthven, I. 2004 A simulated study of implicit feedback models. In *Proceedings of the 26th European Conference on Information retrieval (ECIR 2004)*, Sunderland, UK, Springer Verlag LNCS vol. 2997, pp. 311–326.

Wiebe, J., Hirst, G., and Horton, D. Language use in context. *Communications of the ACM* **39**(1), 102–111.

Wilkinson, R. and Wu, M. 2004 Evaluation experiments and experience from the perspective of interactive information retrieval. In *Proceedings of the 3rd Workshop on Empirical Evaluation of Adaptive Systems, in Conjunction with the 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Eindhoven, The Netherlands, pp. 221–230.

Zadeh, L. 1965 Fuzzy sets. *Information and Control* **8**, 338–353.

https://doi.org/10.1017/S0269888907001282 Published online by Cambridge University Press