Learning Implicit User Interests Using Ontology and Search History for Personalization

Mariam Daoud [1], Lynda Tamine[1], Mohand Boughanem[1], and Bilal Chebaro[2]

[1] IRIT team SIG-RI, IRIT, Toulouse, France
{*daoud*, *tamine*, *bougha*}@irit.fr
[2] Lebanese university, Faculty of Sciences, Beirut, Lebanon
bchebaro@ul.edu.lb

Abstract. The key for providing a robust context for personalized information retrieval is to build a library which gathers the long term and the short term user's interests and then using it in the retrieval process in order to deliver results that better meet the user's information needs. In this paper, we present an enhanced approach for learning a semantic representation of the underlying user's interests using the search history and a predefined ontology. The basic idea is to learn the user's interests by collecting evidence from his search history and represent them conceptually using the concept hierarchy of the ontology. We also involve a dynamic method which tracks changes of the short term user's interests using a correlation metric measure in order to learn and maintain the user's interests.

Key words: user's interests, search history, concept hierarchy, personalized information retrieval

1 Introduction

The explosion of the information available on the Internet and its heterogeneity present a challenge for keyword based search technologies to find useful information for users [2][11]. These technologies have a deterministic behavior in the sense that they return the same set of documents for all the users submitting the same query at a certain time. On the other hand, the effectiveness of these technologies is decreased by the ambiguity of the user's query, the wide spectrum of users and the diversity of their information needs. Recent studies [2] show that the main reason is that they do not take into account the user context in the retrieval process.

The development of relevance feedback [15] and word sense disambiguation techniques [16] aim to assist the user in the formulation of a targeted query, and have shown an improvement of the information retrieval (IR) performance. Effectively, relevance feedback techniques require that a user explicitly provides feedback information, such as marking a subset of retrieved documents as relevant documents. On the other hand, the word sense disambiguation techniques use generally an ontology-based clarification interface and require that the user

specify explicitly the information need. However, since these techniques force the user to provide additional activities, a user may be reluctant to provide such feedback and the effectiveness of these techniques may be limited in real world applications [5].

The above situation gave rise to contextual IR which aims to personalize the IR process by integrating the user context into the retrieval process in order to return personalized results. In [1] contextual IR is defined as follows: *Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs.*

It is common knowledge that several forms of context exist in the area of contextual IR. The cognitive context reflects the user's domains of interest and preferences about the quality of the results returned by the system such as freshness, credibility of the source of the information, etc. We cite also the physical context which reflects constraints on the materials and the user geographical locality, etc. Personalized IR is done when additional information about the user context defined previously is integrated into the IR process so as contextual IR takes place. We present some works within the scope of the personalized IR where the context is modeled as being a user profile representing the user's interests. These works explored various techniques to build the user profile using implicit feedback techniques [10][16][18][3].

In order to endow personalized IR systems with the capability to focus their knowledge on the user's domains of interests, we extend in this paper a related work [17] on building and learning the user's interests across past search sessions in order to enhance the keyword representation of the user's interests to a semantic representation one using a concept hierarchy. We use in our approach both of the search history and the concept hierarchy to learn and maintain the long term user's interests at the time the user conducts a search. We also involve a method which tracks the changes in the short term user's interests.

The paper is organized as follows: Section 2 reviews previous works on personalized IR that learn and maintain the user's interests. Section 3 presents our extended approach of representing and maintaining the user's interests during search sessions. Finally, some conclusions and future works are given in section 4.

2 Related work

Traditional retrieval models and system design are based solely on the query and the document collection which leads to providing the same set of results for different users when the same query is submitted. The limitation of such systems is that the retrieval decision is made out of the search context while the IR takes place in context. Effectively, the IR process depends on time, place, history of interaction, task in hand, and a range of other factors that are not given explicitly but are implicit in the interaction and the ambient environment, namely *the context* [4]. The definition of context in IR is widely abused. While

the wireless networks provide IR possibilities that the users are embedded in a physical environment, a physical context has to be considered in the IR models so as the contextual IR takes place. Personalized IR aims to enhance the retrieval process by integrating the user context or the user profile into the IR process. Works in this area have explored several techniques to build and maintain the user profile using implicit feedback techniques. User's interests are often represented by keyword vectors [7] [17], concept vectors [10] or a concept hierarchy [6][9].

A representation of the user's interests as a concept hierarchy is explained in [6]. An implicit user interest hierarchy (UIH) is learned from a set of web pages visited by the user. A clustering algorithm is applied to group words of the documents into a hierarchy where the high level nodes reflect a more general interest and the leaf nodes are considered more specific and reflect the short term interests.

Webpersonae [10] is a personalized web browsing system based on a user profile that reflects multiple domains of interest. Each one is represented by a cluster of weighted terms. These domains of interest are built by clustering the web pages visited by the user. The system involves the recognition of the current domain of interest used to rerank the search results by comparing the vector representation of recent pages consulted by the user to each of the long term domains of interest.

Recent works exploit ontology-based contextual information to get a semantic representation of the user's interest. Ontology is a concept hierarchy organized with "'is-a"' relationships between them. Many efforts are underway to construct domain specific ontologies that can be used by web content providers. Effectively, the information overload on the Web increases the attempts to provide conceptual search where the semantic web takes place. This research area implies the use of the knowledge representation language [19][14] in order to specify the meaning of the web content according to a concept taxonomy.

ARCH [16] is a personalized IR system that uses both of the user profile which contains several topics of interest and the yahoo concept hierarchy to enhance the user query. The system represents the long term user context as a set of pairs by encapsulating the selected concepts and the deselected concepts that are relevant to the user's information need across search sessions. The short term context is the pair of the selected and the deselected concepts in the current search session. When a long term user context exceeds a similarity threshold with the short term context, the system updates it by combining it with the short term context.

Moreover, Vallet et al. [18] exploit a semantic representation of the user interest based on weighted concept vectors derived from ontology. They build a dynamic semantic representation of the current context which reflects the ongoing user's retrieval tasks and use it to activate a long term user preference or a user's topic of interest. The current context is updated dynamically by using the user's query and feedback information. The personalization is achieved by

re-ranking the search results where the original score is combined with the score yielded by the similarity between the current context and the document.

Challam et al.[3] build a short term user contextual profile as a weighted ontology and use the ODP as reference ontology [13]. The ODP is a Web directory where its purpose is to list and categorize web sites. In this study, the weight of the concept reflects the degree to which it represents the current user's activities. This weight is computed using a classifier to classify a web page into a concept of the ontology. The classification consists on a similarity measure between web page's vector visited by the user and each concept vector representation of the ontology. Thus, the concept's weight is the accumulated weights of all the pages that are classified into the concept and summed with the weights of all children's concepts weights. This user profile is used to re-rank the search results by combining the original rank of the document and the conceptual rank computed using a similarity between the document and the user contextual profile.

This paper presents a new technique for building and learning the user's interests across past search sessions. We exploit in our approach both of the search history and the ODP ontology to learn the long term user's interests at the time the user conducts a search.

Comparatively to previous work in the same area, our approach has the following features:

- A semantic representation of the user context as being a weighted portion of a global ontology with taking into account the short term and the long term user's interests.
- A robust method to detect dynamically related and unrelated user's interests using a statistical rank-order correlation operator between the semantic representation of the user's contexts.

3 Building and maintaining a semantic representation of the user interest

Our main goal is to learn and maintain implicitly the long term user's interests through the passive observation of his behavior. We exploit a cognitive context for our retrieval model where the user's interests are represented semantically. We extend the keyword representation of the user's interests to get an enhanced one using the ODP ontology. In the remainder of this paper we use the term user context as being a vector reflecting the user interest at a certain time.

3.1 Building the user interest

Our method runs in two main steps that are presented in the subsections that follow:

- The first one consists on building the user's interests using an intermediate representation of the user context which is a keyword-based representation in order to get a concept-based representation using the ODP ontology.

– The second step consists on learning and maintaining the user's interests. The learning algorithm is based on a correlation measure used to estimate the level of changes in the semantic representation of the user context during a period of time.

The term-based representation of the user interest: an overview We present in this section an overview of the term-based building process of the user's interests developed in a previous work [17]. The user is modeled by two related components: an aggregative representation of the user search history and a library of user contexts reflecting his interests when seeking information. More precisely, our approach uses the evidence collected across successive search sessions in order to track potential changes in the user's interests. At time s, the user is modeled by $U = (H^s, I^s)$ where H^s and I^s represent respectively the search history and a set of user's interests at time s. A matrix representation is used to represent the search history which is the aggregation of the search session matrix. Let q^s be the query submitted by a specific user U at the retrieval session performed at time s. We assume that a document retrieved by the search engine with respect to q^s is relevant if it is explicitly judged relevant by the user or else, some implicit measures of the user interest such as page dwell time, click through and user activities like saving, printing etc, can be applied to assume the relevancy of a document. Let D^s be the related set of assumed relevant documents for the search session S^s, $R_u^s = \cup_{i=s_0..s} D^i$ represents the potential space search of the user across the past search sessions. We use matrices to represent both user search session and search history. The construction of the search session matrix, described below, is based on the user's search record and some features inferred from the user's relevancy point of view. The user search session is represented by a Document-Term matrix S^s: $D^s * T^s$ where T^s is the set of terms indexing D^s (T^s is a part of all the representative terms of the previous relevant documents, denoted $T(R_u^s)$). Each row in the matrix S^s represents a document $d \in D^s$, each column represents a term $t \in T^s$. In order to improve the accuracy of document-term representation, the approach introduces in the weighting scheme a factor that reflects the user's interests for specific terms. For this purpose, it uses term dependencies as association rules checked among T^s [8] in order to compute the user term relevance value of term t in document d at time s denoted $RTV^s(t, d)$:

$$RTV^s(t,d) = \frac{w_{td}}{dl} * \sum_{t' \neq t, t' \in D^s} cooc(t,t') \tag{1}$$

w_{td} is the common Tf-Idf weight of the term t in the document d, dl is the length of the document d, $cooc(t, t')$ is the confidence value of the rule $(t \rightarrow t')$, $ccoc(t,t') = \frac{n_{tt'}}{n_t * n_{t'}}$, $n_{tt'}$ is the number of documents among D^s containing t and t', n_t is the number of documents among D^s containing t and n'_t is the number of documents among D^s containing t'. $S^s(d, t)$ is then determined as:

$$S^s = RTV^s(t,d) \tag{2}$$

The user search history is a $R_u^s * T(R_u^s)$ matrix, denoted H^s, built dynamically by reporting document information from the matrix S^s and using an aggregative operator combining for each term, its basic term weight and relevance term value computed across the past search sessions as described above. More precisely, the matrix H^s is built as follows:

$H^0(d,t) = S^0(d,t)$

$$H^{s+1}(d,t) = H^s \oplus S^{s+1} = \begin{cases} \alpha * w_{t,d} + \beta * S^{s+1}(d,t) \\ if\ t \notin T(R_u^{(s)}) \\ \alpha * H^s(d,t) + \beta * S^{s+1}(d,t) \\ if\ t \in T(R_u^{(s)}) \\ H^s(d,t)\ otherwise \end{cases} \tag{3}$$

$(\alpha + \beta = 1), s > s_0$

After the representation of the search history, a weighted keyword representation of the user context K^s is extracted and reflects the user's interests at learning time s. The term's weight reflects the degree to which the term represents the user context. It is computed by summing for each term in $T(R_u^s)$ the columns in H^s as follows:

$$c^s(t) = \sum_{d \in R_u^s} H^s(d,t) \tag{4}$$

$K^s(t)$ is normalized as follows: $c^s(t) = \frac{K^s(t)}{\sum_{t \in T^s} K^s(t)}$. This original approach models the user context as a set of weighted keywords reflecting the user's interests in a search session. Given a user being interested in the military domain for a given search session, then we find terms of the military domain in the top of the term-based representation of the user context. The maintaining process of the user's interests between search sessions is accomplished using a rank order correlation measure applied on two consecutive keyword contexts of retrieval session. Change to a different domain interest contribute to add or change the keywords of the user context. This approach is faceted to a risk error due to the lack of not taking into account the semantic relation between words, so as the changes of interests depend on a distinctive difference in the rank order distribution of the keyword-based context representation, independently of their belonging to the same user's information need. We aim to enhance the keyword representation of the user interest to a semantic representation one that outcomes the limit of tracking changes in the user's contexts. Related and unrelated user's contexts are detected using the same measure but applied on a semantic representation of them. Effectively, enriching the keyword representation of the user's interests with concepts from the core ontology has two benefits: first, instead of the keyword representation, it provides a semantic meaning of the user's interests. Second, tracking the changes of the short term user interest is more reliable and accurate when they are represented semantically.

The concept-based representation of the user interest using ontology We present in this section the method for a concept-based representation of the user interest in a semantic context to be stored in I^s. To get the semantic representation of the user context, we map the keyword vector representation of the user context described in the previous section on the concepts of the ODP ontology [13], thus we obtain a weighted concept hierarchy which is the semantic representation of the user context at a certain time.

– *Reference ontology and representation of domain knowledge* There are many subject hierarchies created manually and designed to organize web content for easy browsing by end users. We cite the online portals such as yahoo [1], Magellan[2], Lycos [3] and the open directory project [13]. Considering that the Open Directory Project (ODP) is the most widely distributed data base of Web content classified by humans, we use it in our profiling component as a fundamental source of a semantic knowledge to represent semantically the user's interests. We show in Fig.1. the concept hierarchy of the ODP ontology. Various methods can be utilized to represent the concept vector of the ODP

Fig. 1. The concepts in the ODP ontology

ontology. In our approach, we use a term-vector based representation for the concepts developed in [3]. We are interested by the top three levels of the ontology to represent a set of general user's interests. Each concept of the hierarchy is associated to a set of related web pages. These documents

[1] http://www.yahoo.com
[2] http://www.mckinley.com
[3] http://point.lycos.com/categories/index.html

are used to represent the term vector representation of the concept. The content of the pages associated to the concept j are merged together to create a super-document sd_j to obtain a collection of super-documents, one per concept, that are pre-processed to remove stop words and stemmed using the porter stemmer to remove common suffixes. Thus, each concept is treated as a n-dimensional vector in which n represents the number of unique terms in the vocabulary. Each term's weight in the concept's vector is computed using the $tf*idf$ weighting scheme and normalized by their length. The term weight of the term i in concept j is computed as follows:

$$w_{ij} = tf_{ij} * idf_i \tag{5}$$

Where
tf_{ij}=number of occurrences of t_i in sd_j
N=the number of super-documents in the collection
n_i=the number of super-documents containing t_i

– *Semantic representation of the user interest* The basic idea to get a semantic representation of the user context is to map the keyword representation of the user context on the concept hierarchy. Concepts and user context are represented in the vector space model as explained in the previous section. Thus the mapping consists on a cosine similarity between vectors and has as output a weighted concept vector which represents the semantic representation of the user context. The semantic vector c^s represents the short term user context at learning time s and includes his short term interests. The dimension of the semantic vector c^s at learning time s is equal to the dimension of the top three levels of the ODP's domain ontology θ. The weight of a concept in the ontology reflects the degree to which it represents the user's short term interest and beliefs at time learning s. Let K^s be the keyword representation of the user context computed as explained in the previous section and V_j the term vector representation of a concept j from the ontology. The concept's weight is then computed as follows:

$$P_j = cos(V_j, K^s) \tag{6}$$

We then order the concept vector by decreasing weight where concepts with high weights reflect the short term user interest.
These semantic user's interests are then reused in the various phases of the personalized information access. Thus, the library of the user's interests can be used for the:

- Query reformulation
- Query to document matching
- Re-ranking the search results

As example, given a user being interested by the field of computers in a certain search session. *Computers* is categorized at the first level of the ODP's concept hierarchy. We take into account the top three levels of the concept hierarchy, we suppose a more specific user interest in the search session to be

the *software* which is the subcategory of *computers* and *malicious software* which is the subcategory of *software* category as shown is Fig.2.
We assume that the keyword representation of the user context include terms related to the specific domain of interest cited above, and are extracted from the user search history. Thus, certainly the keyword-based user context, mapped to the nodes of the concept hierarchy exceeds a similarity threshold with the vector representation of the category *computers* and its subcategory *software*, especially the *malicious software*. In this way, we identify the best matching concepts for the keyword user context. The semantic vector representation of the user context is then generated on the basis of concept's weights and the categories cited above, having the high weights, are ordered in the top of the vector-based semantic user context.

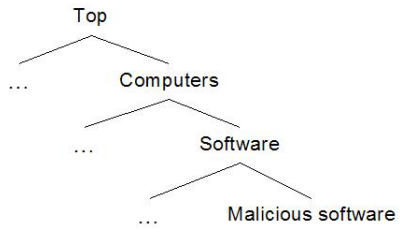

Fig. 2. A personal user's interests in a search session

3.2 Maintaining the user's interests

As long as the user conducts a search, the system must tracks the changes in the short term user context. A new user context means a new user interest must be added to the library of the user's interests or a long term user interest may be reviewed. The key for providing an accurate profiling of the user's interests is to determine at what degree we shall update an existing user interest and when we shall add a new user interest to the library of the user's contexts. In our approach, we compare the current semantic context at time s noted cc^s and the previous one pc^s using Kendall rank-order correlation operator as showed in Fig.3. The Kendall rank correlation coefficient evaluates the degree of similarity between two sets of ranks given to a same set of objects. In our case the objects are the concepts of the ontology representing the user's interests. A change of domain of

interest between search sessions contribute to change or to add keywords in the search history matrix, then to change the keyword based representation of the user context, we then conclude a significant change of the rank order of concepts being in the top of the concept-based representation of the current user's context. The Kendall rank correlation coefficient is given in the following formula:

$$\Delta I = (cc^s \circ pc^s) = \sum_{o \in \theta} (cc^s(o) - pc^s(o)) \tag{7}$$

where θ is the set of the top three levels of the ODP's concept hierarchy. The coefficient value ΔI is in the range [-1 1], where a value closer to -1 means that the semantic contexts are not similar and a value closer to 1 means that the semantic contexts are very related to each other. Based on this coefficient value, we apply the following strategy in order to learn the user's interests and so update the set of user interests in I^s:

1. $\Delta I > \sigma$ (σ represents a threshold correlation value). No potential changes in the user's contexts, no information available to update I^s;
2. $\Delta I < \sigma$. There is a change in the user's contexts. In this case we gauge the level of change, and two configurations may be presented: the change implies a refinement of a prior detected user's interest or else the occurrence of a novel one. In order to answer this question we do as follows:
   - select $c^* = argmax_{c \in I^s}(c \circ cc^s)$,
   - if $cc^s \circ c^* > \sigma$ then
     - refine the user's interest c^*: we define a refinement formula that combine the newly constructed semantic vector with the user interest c^* where the concepts weights computed in c^* are automatically reduced by a decay factor ζ, a real value in $[0, 1]$. The refinement of c^* is given as follows:
       $c^* = \zeta * c^* + (1 - \zeta) * cc$;
     - update the matrix H^s by dropping the rows representing the least recently documents updated, update consequently R^s_u,
   - if $cc^s \circ c^* < \sigma$ then add the new tracked interest in the library I^s, try to learn c^* a period of time by reinitializing the search history matrix to be equal to the current search session matrix as follows:
     set $H^{s+1} = S^s$, $s_0 = s$

An updating procedure of the library of the user's interests consists on managing their persistence. Indeed, it consists on removing some user's interests according to the updating frequency or the date of the last update. By this fact, we exclude the non recurrent contexts inserted in the library of the user's interests.

4 Conclusion and future works

We proposed in this paper a new approach for building an ontology-based user's interests in the field of personalized IR. We improved a previous work for user

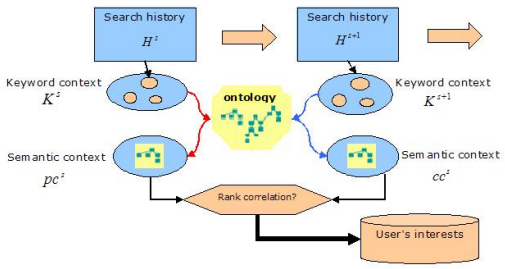

Fig. 3. Learning and maintaining process of the user's interests

modeling where the user interest consists on a keyword-based representation. In order to enhance the original approach, we exploit both the search history and a predefined ontology ODP to represent the user's interests conceptually. The basic idea consists on mapping the keyword representation of the user context to the concept hierarchy, and then each concept has a weight that reflects the degree to which it represents the user context at a certain time. The approach integrates the temporal dimension in the user's interests learning process. More precisely, we learn and maintain long term user's interests by updating the search history representation using the user relevancy point of view on familiar words from which we extract a short term user's interest.

A distinctive aspect in our approach is the use of the kendall rank order correlation measure between semantic representations of the user's interests. The benefit of using this measure is the gain of accuracy and reliability in tracking changes of the short term user's interests instead of applying it on a keyword-based representation of the user interests.

In future work, we plan to improve the maintaining process of the short term user's interests; Instead of tracking the changes of the user context by the user's queries submitted among search sessions, we aim to integrate a method for detecting session boundaries in order to activate the method for updating the library of the user's contexts. We define a session as a set of queries related to the same information need. We also tend to personalize the information retrieval process by using the user's interests in the query reformulation. We use the term of the concept representing the short term user interest in order to enhance the user query and personalize the search results to better meet the user's information needs in a search session.

In another hand, we plan to evaluate our approach experimentally using a large scale of quantitative data on the user search sessions and accurate contexts provided by the related queries during a reasonable period of testing a particular search engine.

References

1. Allan, J., al.: Challenges in information retrieval and langage modelling. In: Workshop held at the center for intelligent information retrieval, Septembre (2002)
2. Budzik, J., Hammond, K.J.: Users interactions with everyday applications as context for just-in-time information access. In: Proceedings of the 5th international conference on intelligent user interfaces,(2000) 41–51
3. Challam, V., Gauch, S., Chandramouli, A.: Contextual Search Using Ontology-Based User Profiles. In: Proceedings of RIAO 2007, Pittsburgh USA 30 may - 1 june (2007)
4. Ingwersen, P.,Jarvelin, K.: Information Retrieval in Context IRiX. In: ACM SIGIR forum,(2005)
5. Kelly, D.,Teevan, J.: mplicit feedback for inferring user preference: A bibliography. In: SIGIR Forum,(2003)
6. Kim, H. R., Chan, P. K.: Learning implicit user interest hierarchy for context in personalization. In: Proceedings of the 8th international Conference on intelligent User interfaces IUI '03, Miami Florida USA January 12 - 15 (2003)
7. Lieberman, H.: Letizia:An agent thatassists web browsing. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI95), Montreal August (1995) 924–929
8. Lin, S.H., Shih, C.S.,Chen, M.C.,Ho, J.,Ko, M.,Huang, Y. M.: Extracting classification knowledge of Internet documents with mining term-associations: A semantic approach. In: the 21th International SIGIR Conference on Research end Development in Information Retrieval, (1998)
9. Liu,F., Yu, C., Meng, W.: Personalized Web Search For Improving Retrieval Effectiveness. In: IEEE Transactions on Knowledge and Data Engineering,Vol. 16(1)., January (2004) 28–40
10. Mc Gowan, J.P.: A multiple model approach to personalised information access. In: Master Thesis in computer science, Faculty of science, University College Dublin, February (2003)
11. Nunberg G.: As Google goes, so goes the nation, New York times, May 2003
12. M. Pazzani, D. Billsus, "Learning and revising user profiles: The identification of interesting Web sites", Machine learning, Vol 27, pp 313-331, 1997
13. The Open Directory Project (ODP), `http://www.dmoz.org`
14. Fensel, D., Harmelen,F.V., Horrocks, I., Deborah: OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems, Vol. 16, No. 2, March/April 2001.
15. Rocchio, J.: Relevance feedback in information retrieval. In: Salton, G. (ed.): The SMART retrieval system - experiments in automated document processing. Prentice-Hall, Englewood Cliffs, NJ (1971)
16. Sieg, A., Mobasher, B., Burke, R.,Prabu, G., Lytinen, S.: representing user information context with ontologies.
17. Tamine, L., Boughanem, M., Zemirli, W.N.: Inferring the user's interests using the search history. In:Workshop on information retrieval, Learning, Knowledge and Adaptability (LWA 2006), ildesheim Germany november 9 - 11 (2006)108–110
18. Vallet, D., Fernndez, M., Castells, P., Mylonas, Ph., Avrithis, Y.: Personalized Information Retrieval in Context. In: 3rd International Workshop on Modeling and Retrieval of Context, Boston USA 16-17 July (2006)
19. Lassila, O., Swick, R.: Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium recommendation, 22 February (1999)