2009-06-16

A context-aware approach based on self-organizing maps to study web-users' tendencies from their behaviour

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Stephen Barrett
Trinity College Dublin, Ireland

Follow this and additional works at: https://arrow.tudublin.ie/scschcomcon

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation

A Context-Aware Approach Based on Self-organizing maps to study Web-Users' tendencies from their behaviour

Luca Longo
Department of Computer Science
and Statistics
Trinity College Dublin - Ireland
llongo@cs.tcd.ie

Stephen Barrett
Department of Computer Science
and Statistics
Trinity College Dublin - Ireland
stephen.barrett@cs.tcd.ie

ABSTRACT

In the context of a highly volatile web of uneven quality, the identification of content deemed valuable by end users is of paramount importance. Where page content undergoes rapid change, this issue is particularly challenging. Web browsing activity represents a unique source of context by which the value of web pages can be determined via an assessment of individual user interactions, such as scrolling, clicking, saving and so forth. Over time, this data set forms a pattern of activity which can be mined for meaning. In this paper we present an approach to web content, based on Kohonen mapping, used to generate a topological model of users' behaviour over web-pages. Each web-document can thus be represented as a semantic map built by adopting unsupervised techniques where similar users' behaviour are mapped close together, with identification of information stability emerging as a by product of the identification of similarity in user activity over content. In this model, the more similar the outputs of the map for each user who has endorsed a web-page, the more the web site is considered current or in context with changing information. We illustrate the potential application of this approach to our ongoing work in social search.

Categories and Subject Descriptors
L.2.1 [**Computing Methodologies**]: Artificial Intelligence
—*Distributed Artificial Intelligence*

General Terms
Context-awareness, Computational Trust, Distributed Systems, Social Search, Self-Organizing map

Keywords
Web, Social Search, Kohonen maps

1. INTRODUCTION

Human behaviour is influenced by the context in which we live. In response to context, we act typically in common ways, within a predictable range, even if atypical response is sometime apparent. In society, users' behaviour is influenced by various categorising factors, such as the context in which they interact, the experience they have in that context, and their knowledge, experience and goals. We are expert in categorising people in practice, implicitly utilising many of these contextual cues. In an analogous way, in the World Wide Web it is possible to discern different kinds of users. The same behavioural cues can be identified in web browsing activity, in the footprints we leave in our forms of interaction, captured in the set of our actions, such as scrolling, clicking, cut and paste, bookmarking, printing and so forth. Users can thus be categorised in theory, for example, from the more expert who can find information straight away to the novice who requires more effort to discover information. Unfortunately, in practice it is difficult to discriminate specific users as behaviour is difficult to distinguish amid the scale of data.

In this paper, we propose a methodology, based on a popular unsupervised technique called *self-organizing maps*, to computationally discriminate users by considering the digital footprint generated by their online browsing behaviour. We deploy Kohonen maps to create a topological structure describing users' behaviour over web-pages. Via this technique, we hypothesise that we can interpret a common region of a Kohonen map amongst a set of users who have viewed the same web document to indicate that the content of a web page is stable, suggesting a strong degree of trustworthiness as to a proposition. For instance, if the emerged common behaviour belongs to a region of the Kohonen map that describes behavioural indifference regarding a web page, we can support the proposition that the considered web-page carries futile and unprofitable information. Similarly, if the region describes behavioural interest in the page, we can confidently suppose that the given web-document contain useful and advantageous data. If a common behavioural response does not emerge from the topological structure of the map, no justification of the trustworthiness of a web entity can be extracted. This evidence can be interpreted as a web-site instability as any advantage or benefit can be achieved.
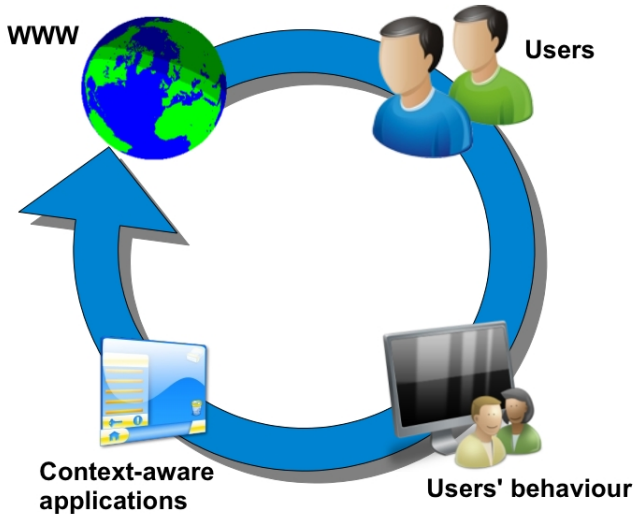
Figure 1: The circular behavioural context-aware process

With the above considerations, we can adapt the context via applications by taking users' behaviour into account. In particular, the automatic addition of new behaviour and context acquired from users interaction within the WWW, facilitates online learning, enabling applications to evolve with changing user requirements, preferences or tendencies. The main goal of our proposal is to capture context via web-user's behaviour to be used in context-aware applications such as search-engines, information retrieval middleware, and context-aware frameworks. These applications influence users' behaviour within the context, closing the circle as depicted in figure 1

The particular contributions of this paper are organised as follows. In section 2 we situate our work in the area of context awareness, social search, and introduce the concept of the self-organizing map. In section 3 we describes how we model users' behaviour over web-pages, and present our theoretical model along with our hypothesis. We conclude in section 4 by outlining open issues and future works.

2. RELATED WORK

2.1 Context Awareness

Context awareness deals with synchronising application behaviour with changes in a dynamic environment. There are several definitions of *context* in the literature relevant to this. Chen and Kotz [2] differentiate active and passive aspects of context by defining it as *a set of environmental states and settings that either determines an application's behaviour or in which an application event occurs and is interesting to the user.* Schilit and Theimer [3] refer to context as *location, identities of nearby people and objects, and changes to those objects.* A more general definition by Dey et al. [4] describes context as *any information that can be used to characterise the situation of an entity. An entity is a person,*

place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves.

Recent research in context-awarness computing focus on user-related properties to try to better develop applications by considering users' preferences, tendencies, their knowledge and behaviour. In [19] and [20] both authors demonstrated how users' properties can be successfully integrated with other context-related functionality, by taking into account the current context of the user. Byun and Cheverst [21] made use of historic context and behaviour for aiding adaptation. They investigated how context history could aid a system in exhibiting intelligent and proactive behaviour.

The issue of context is becoming important for web-based researchers with the advent of hybrid search engines that seek to rank web search results in a manner informed by human browsing activity. This type of ranking system differs from traditional engines, such as Google [15], that relies on the link structure of the web to predict users' tendencies. Hybrid search engines must compute over significantly more complex meta-data regarding human factors, and are considered to be an important development for increasing the relevance and performance of information search [1]. This work can be characterised as an effort to modify ranking behaviour to account for end user context, both in terms perhaps of personalisation of results to account for individual preference as identified by previous browsing actions, and in terms of the determination of content value via analysis of end user browsing activity footprint.

2.2 Social Search

The World Wide Web has been growing rapidly in size and web-search technologies have become integral part of it. *Social search* is a web-search approach based on the activity of contributions to assign the relevance of web-search results. Several computational methods have been conceived combining human intelligence with computer paradigms [6] [7] and implicit feedback strategies allow the automatic evaluation of web-entities that users show interest in. In [11] an HTTP tool was developed, logging data about mouse movements, keyboard input and more. Here, authors focused on tasks such as classifying the user with regard to computer usage proficiency or making a detailed assessment of how long it took users to fill in fields of a form. Similarly in the work of Velayathan and Yamada [12], an unobtrusive framework has been proposed to gather logs of users' behaviour. These logs are then analysed in order to extract effective rules to evaluate web-pages using a machine-learning techniques. In [13] three sources of implicit feedback are taken into account: reading time per document, scrolling and interaction, focused on the hypothesis that users will spend more time, scroll more often and interact more with those documents they find relevant. There are further several works in the literature, but very few try to take into consideration the full range of possible events that a web browser can capture.

2.3 Self-organizing maps

The Kohonen neural network [8] is a very common architecture, also known as self-organizing maps (SOM), that uses unsupervised techniques to learn and discover some underlying structure of a data set. Kohonen's SOM is a topology-preserving map because it imposes a topological structure on the nodes in the network that preserves neighbourhood relations. The model proposed by Kohonen differs from the feedforward backpropagation network [10] in different ways. The network is trained in an unsupervised way so data is given in input but there is no anticipated output. Training builds the topological structure using input examples, by finding which output neuron wins. This winning neuron's weights are the modified so that it will activate higher on the pattern that caused it to win. Finally, mapping automatically classifies a new input vector. The goal of learning in a Kohonen map is to cause region of the network to respond similarly to certain input patterns. If the membership of a specific class is know a priori, given an input, then it is possible to associate each region to the related class. This property is extremely interesting for our work, as it allows us to classify similar data capturing distinct user's behaviour into classes. So, if we know that a user's behaviour pattern belongs, for instance, to the class of *positive behaviour*, we are able to map similar patterns into the same class. Furthermore, if a common users' behaviour emerge, given a web-document, we are able to rate trustworthiness of the proposition that some particular behavioural response is correct for a user group, and so classify a document in a manner that allows us to predict users' responses.

The Kohonen map paradigm has inspired various relevant research. In [16] the authors made use of the cluster information generated by a self organizing-map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions. Data captured includes traffic volume, types of files accessed, domain summary, operating system used, navigation and soon. In [17] the authors take into consideration hyperlinks as an indication of a member interest. They applied their method based on Kohonen mapping into the Blogalia weblog community in order to discover community features. Velasquez et al. [18] adopted the self-organizing approach to identify clusters of similar web-sessions for the analysis of visitor behaviour. Two pieces of evidence were considered to define the visitor's behaviour: the visited pages and the time spent in each one of them. In [22] the authors propose a method based on a prior clustering of web-pages that uses semantic information, such as users internet sessions, at the server-side. Kaski et al. [23] presented a framework based solely on the web-site content whereas in [24] date, time and address of the requested web pages, as well as the Ip address of the user's machine, are taken into account to form a relevant pattern for the Self-organizing map training process. The approach proposed in this paper differs from the ones described above as the authors consider the more psychological side of the concept of behaviour by analysing more actions, performed by users, upon web-pages. Along with the time spent on a web document, more pieces of evidence can be considered such as scrolling activity, highlighting text as described in the following section.

3. HYPOTHESIS AND THE THEORETICAL MODEL

During an Internet browsing session, users interact with the browser by performing a set of actions that reflects his/her behaviour. Of the various possible actions to be analysed, early research suggests that the following events represent possible pieces of evidence useful for computation of user value judgements regarding web-documents [5].

- bookmarking;
- printing;
- saving as (page);
- downloading;
- submitting (text form filling);
- saving as (picture);
- cutting & pasting (clipboard using);
- scrolling;
- finding in page;
- highlighting text.

Furthermore, the authors believe that temporal-based evidence should also be considered in order to determine the relevance of a digital entity [9]. For instance, a user's inactivity period and focused time on a particular web-document can be easily calculated. The set of all these actions form a relevant pattern of data that we refer to as a *user's behaviour pattern*.

The authors of this paper have performed various experiments in evidence gathering and processing in order to evaluate a web-page automatically, making use of a robust anonymous client-side logging tool designed for the purpose, that unobtrusively gathers logs of users' behaviour. Previous experiments they conducted were focused on the analysis of participants' behaviour while performing submitted task [5]. In particular, they analysed users' effort spent on a set of web-documents to complete a proposed task. Volunteers were asked to explicitly rate the usefulness of the information gathered in each web-document in order to perform a comparison between implicitly obtained result. The product of the comparison showed an encouraging relationship between implicit feedback and explicit judgements, that suggests the feasibility of replacing explicit human ratings with implicitly computed rankings. At this point, they were able to predict user tendencies by proposing a base ranking system of web documents. However, the number of participants involved in previous experiments was not large (of the order of a hundred), so the results were tentative. By using the self-organizing map paradigm we intend to analyse the behaviour patterns, of large numbers of users, as gathered by our logging tool, and to extract and to cluster common behaviours without relying on the qualitative analysis of each pattern's content. New learning patterns of behaviours can be added to the map at any time and will be automatically organised into the proper cells of the map, hence obtaining information

more current and in context. The simplicity of adding new resources to the system distinguishes map-based navigation from the concept-based approach, that previously required a significant investment of time for manual indexing. This is an important issue, as it underlines the degree of flexibility and adaptation of our approach that echoes to the evolution of the World Wide Web.

By taking into account the fact that several web-users' behaviour can be encountered during the logging activity, and since there is no output pair related to them, the authors consider the self-organizing map paradigm to be appropriate in identifying similar habits, tendencies and affinities among web-users. In this case the relevant information has to be found within the training samples, i.e. the set of *user's behaviour patterns*.

We plan to do experiments with users by implicitly monitoring their behaviour while surfing the Web. Furthermore, we will ask experts to explicitly rate experiment volunteers' human behaviour during browsing session, without analysing the digital pattern, but simply evaluating their human cognitive behaviour, in order to build up a set of pre-defined common behaviour classes that can be used as labels for the outputs of the Kohonen model. For instance, we would like to create generic classes denoting good behaviour, bad behaviour, quick behaviour, focused behaviour and so forth. The expertise will be a gate between explicit evaluation and implicit mapping. The Kohonen neural network training will be done by submitting to it all the *user's behaviour pattern* logs gathered by our logger. Subsequently, an analysis of the outputs, will be performed trying to associate labels, previously estimated by considering expert's opinion, to common regions of the topological map as they emerge from the neural network.

The hypothesis behind this work relies on the analysis of the outputs of the self-organizing map, following up all the new behaviours gathered from users on a given web-document. In particular, the authors are interested in the number of *user's behaviour patterns* that activate the same region of the computed topological structure of the Kohonen map. If the number will be significant, we can be confident of the web-site's content trustworthiness as the emerged common behaviour result to be stable. The calculated trustworthiness can be either positive or negative, but if the majority of users support this evidence, by behaving in the same way, this can be used as a good predictor for web entities. As a consequence of these considerations, a relevant social search engine, based on implicitly monitoring users' behaviour while surfing the Internet, can be conceived. The result can be confidently used in conjunction of a Computational Trust model in supporting its ranking system. In the figure 2 the schema of the Self-Organizing map (S.O.M.) training process and the web-entity evaluation process are depicted. Both the process are summarised in the following algorithms as well.

Self-Organizing map Training process

```
1 MONITOR:
    the logger captures users' behaviour
    over web pages
2 PRE-PROCESSING:
    noisy information is discarded and
    aggregation operations are performed
3 FEATURES EXTRACTION:
    features, i.e. browser's events,
    are considered, for each web page,
    to form the User's Behaviour Pattern
4 SOM TRAINING:
    the entire set of User's Behaviour
    patterns is used to training the Kohonen Map.
    This step may be repeated as soon as a
    relevant amount of new patterns are available.
5 TOPOLOGICAL STRUCTURE:
    the map emerges after  iterating different
    times on the inputs in 4
6 HUMAN EXPERTISE:
    human specialists help in labelling emerged
    map's regions
```

The logger captures all the actions performed by a user on his/her browser. This is a continuos process as it is part of the browsing session: every time a user surfs a web-page, he/she generates a set of events. This step is repeated for each user worldwide who has previously installed our logging tool. The information gathered need to be cleaned from the noisy data, aggregated and pre-processed, in order to extract meaningful features and to produce a valid and appropriate *user's behaviour pattern*, related to a particular user and a given web-entity. The set of the extracted behaviour patterns are periodically used to train the Self-Organizing map in order to produce an up to date topology-preserving map. Expertise is required to associate to the emerged Kohonen map's regions the appropriate labels, i.e. to identify and to give names to output behaviours' regions.

Web-page's Trustworthiness estimation process

```
1 WEB-PAGE:
    a web-document X is taken into account
2 BEHAVIOURS COLLECTOR:
    all the users' behaviour on X are collected,
    i.e. the set of all the User's Behaviour
    Pattern gathered by the Monitor
3 TOPOLOGICAL STRUCTURE:
    all the patterns in 2 are presented,
    one by one, to the neural network
4 CLUSTERS/LABELS:
    each pattern in 3 activate a region of the
    Kohonen map and the related label
    is returned as output by using a formal metric
5 COMPUTATIONAL TRUST MODEL:
    a formal model evaluates the outputs of 4
    labels in order to compute the
    Trustworthiness of web-document X
6 TRUSTWORTHINESS VALUE:
    a trustworthiness value is inferred to
    web-site X by using a specific metric
```
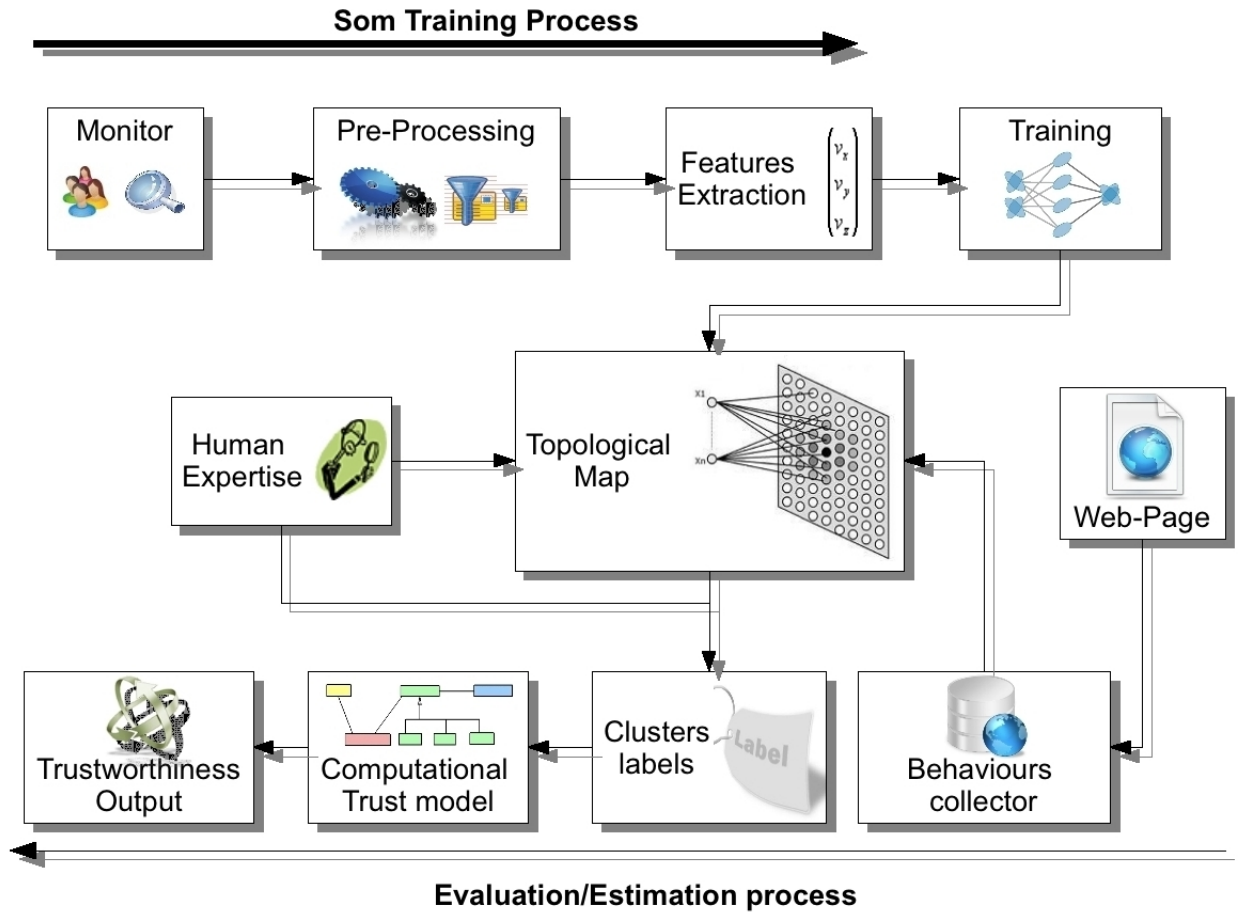
Som Training Process

Monitor → Pre-Processing → Features Extraction $\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}$ → Training

Human Expertise → Topological Map

Web-Page

Trustworthiness Output ← Computational Trust model ← Clusters labels ← Behaviours collector

Evaluation/Estimation process

Figure 2: Training and Evaluation processes

To compute the trustworthiness value of a certain web-page, all the *users' behaviour patterns* related to it, i.e. generated by all the users who previously have viewed that website, have to be forwarded to the Kohonen neural network to retrieve the corresponding output. The Computational Trust model collects, one by one, all the outputs. With an appropriate function, it produces a value indicating the stability of that web-site hence, how much we can be confident of its content. A basic Computational Trust model can return a value in the range of [0..1] indicating the percentage of users whose behaviour has been classified into the same region/cluster of the Kohonen map. More sophisticated models can adopt statistical approaches as, for instance, by using standard deviations and statistical tests to find the true mean of emerged behaviour. Furthermore, the set of labels emerged from the Kohonen map may be used as a supporting-level for a different Trust model: in other words they may be used as a weight for sustaining another Trust model's output. For example, a Trust model may be based on a *cost function* that rates user's effort spent on a web-entity, over time. This value may be high or low: high values indicate good effort underlying a good degree of interest while low values support disinterest.

By using the values emerged from the Self-Organizing map, synonymous of stability, it would be possible to weight the *cost function* to better evaluate sophisticated web-documents. At this stage, a third-generation *social search engine*, in which web-entities are ordered according to their computed trustworthiness degree, may be conceived and developed. This approach represents an interesting challenge both for its robustness, in term of trustworthy outputs, and for its feasibility.

4. CONCLUSION AND OPEN ISSUES

In the context of a highly volatile web of uneven quality, web browsing activity represents a unique source of context for the computation of content deemed valuable by end users. In this position paper, we proposed an approach to assessment of valuation of web content, in terms of stability, trustworthiness and reliability, based on the popular unsupervised Kohonen self-organizing map, used to generate a topological-preserving model of users' behaviour over Internet web-pages. The work describes research on a mining approach to adaptation along with the concepts and steps to build up the model proposed. Artificial neural network models have particular properties such as the ability to adapt, to learn or to cluster data. These models are inspired by our

present understanding of the biological neural system and their goal is to model the humans' brain neural structure with non-linear computational elements interconnected to each other. Each connection is characterised by a variable weight that is adjusted during the so-called learning stage. Self-organizing networks, and in particular the Self- Organizing Feature Map (SOM, sometimes referred as Kohonen maps), are a popular and tested [14] kind of artificial neural network that try to build a representation of some feature of the input vector used in the training-process. This paradigm organises neurons in a lattice, usually by means of a one or two-dimensional array, placed in the input space and is spanned over the input vector distribution. By using a two-dimensional SOM network it is possible to obtain a map of input space where closeness between units or clusters in the map represent closeness of the input vectors.

In this work, the authors adopt the Som's concepts to create a topological structure of the user's behaviour while surfing the Web. Their hypothesis is that, the more similar the outputs of the map for each user who has endorsed a web-page, the more the web site is considered current or in context with changing information. This evidence would support the concept of stability of web-entities and may be adopted with a formal Computational Trust model, to estimate the degree of trustworthiness of web-pages. The aggregation of these concepts would enable us to conceive a third-generation *social search engine* based on users' behaviour. Users' habits, preferences and tendencies may be predicted and Web-entities may be ranked according their computed trustworthiness value. The implementation of the formal model will be made in conjunction of the logging tool testing. The distributed structure of the Web generates communication issues between loggers and the server along with privacy concerns. In order to make our approach feasible, we need to guarantee web-users' anonymity and to adopt robust distributed algorithms to exchange and store data within the network in a safe way.

5. REFERENCES

[1] Schmidt S., Beigl M., Gellersen H-W., *There is more to Context than Location.* Computers & Graphics Journal, Elsevier, 23 (6): 893-902. December 1999.

[2] Chen G., Kotz D. *A survey of context-aware mobile computing research*, tech. report TR2000-381, Dept. of computer Science, Dartmouth College, 2001

[3] Schilit, B., Theimer, M. *Disseminating Active Map Information to Mobile Hosts.* IEEE Network, 8(5). 1994. pp 22-32.

[4] Dey A.K., Abowd G.D., *Towards a Better Understanding of Context and Context-Awareness.* CHI 2000, Workshop on the What, Who, Where, When, and How of Context-Awareness, 2000.

[5] Longo L., Barrett S., Dondio P. *Toward Social Search: from Explicit to Implicit Collaboration to Predict Users' Interests.* ACM WEBIST 2009. 5th International Conference of Web Information Systems and Technologies. Lisboa, Portugal.

[6] Agichtein E., Brill E., Dumais S. *Improving Web Search Ranking by Incorporating User Behavior Information.* SIGIR '06, August 6-11, 2006, Seattle, USA.

[7] Agichtein E., Zheng Z., *Identifying Best Bet Web Search Results by Mining Past User Behavior.* Kdd '06, August 20-23, 2006, Philadelphia, Pennsylvania, USA.

[8] Kohonen, T., *Self-Organization and Associative Memory.* New York: Springer-Verlag, 1988.

[9] Longo L., Dondio P., Barrett S. *Temporal Factors to evaluate trustworthiness of virtual identities.* IEEE, SECURECOMM 2007, Nice, France.

[10] Heaton J.T., *Introduction to Neural Networks with Java* Heaton Research Inc, Editor Mary McKinnis, 2005.

[11] Atterer R. et al., *Knowing the User's Every Move - User Activity Tracking for Website Usability Evaluation and Implicit Interaction.* WWW 2006, May 23-26, Edinburgh.

[12] Velayathan G., Yamada S., *Behavior-based Web Page Evaluation,* WWW 2007, May 8-12, 2007, Banff, Alberta, Canada.

[13] Kelly D. et al., *Reading Time, Scrolling and Interaction: exploring Implicit Sources of User Preferences for Relevance Feedback During Interactive Information Retrieval.* SIGIR' 01, New Orleans, USA.

[14] Kohonen, T., Kaski, S., Lagus, K., and Honkela, T. (1996) "Very Large Two-Level SOM for the Browsing of the Newsgroups". In Proc. ICANN'96 (Berlin:Springer-Verlag), pp. 269-274

[15] Page L. Brin S. Motwani R. and Winograd T. *The PageRank citation ranking: Bringing order to the Web.* 1999. Standford University.

[16] Wanga X., Abrahamb A., A. Smitha K., *Intelligent web traffic mining and analysis.* Journal of Network and Computer Applications 28, pp. 147-165, 2005

[17] Merelo-Guervos J. J., Prieto B., Prieto A., Romeo, G., Castillo Valdivieso P., *Clustering web-based communities Using Self-organizing maps.* IADIS International Conference Web Based Communities, 2004. Lisbon, Portugal, 24-26 March 2004.

[18] Velasquez J.D., Yasuda H., Aoki T., Weber R., Vera E. *Using Self Organizing Feature Maps to acquire knowledge about visitor behavior in a web site* Journal of Lecture Notes in Artificial Intelligence. 2003.

[19] N.A. Bradley and M. D. Dunlop. *Toward a multidisciplinary model of context to support context-aware computing.* Journal of Human-Computer Interaction, 20(4): pp.403-446, 2005.

[20] A. Jameson. *Modeling both the context and the user.* Personal Technologies, 5(1):29-33, 2001.

[21] H.E. Byun and K. Cheverst. *Harnessing context to support proactive behaviors.* In AI in Mobile Systems, France, 2002.

[22] Rossi F., El Golli A., Lechevallier Y., *Usage Guided Clustering of Web Pages with the Median Self Organizing Map.* Proceedings of ESANN '05, pp. 351-356

[23] Kaski S., Hinkela T., Lagus K., Kohonen T. *WEBSOM - self-organizing map of document collections.* Neurocomputing 21, 1-3, October 1998, pp. 101-118.

[24] Smith K.,N. Alan, *Web Page Clustering using a Self-organizing map of User Navigation Patterns.* Decision Support Systems, pp. 35(2):245?256, 2003.