Crowdsourcing Supported Context Detection for Improving Information Search Activities

Michael Beul, Stefan Eicker
paluno - The Ruhr Institute for Software Technology
University of Duisburg-Essen
Essen, Germany
{michael.beul | stefan.eicker}@paluno.uni-due.de

Abstract— **In environments that require the use of software applications, intervals where application functionality, tools, methods and technical systems are changing are often very short. The process of searching for relevant information about a specific issue is frequently executed and time-consuming. Because of the availability of a nearly unlimited amount of data, people spend a lot of time in formulating search queries and evaluating the relevance of the search results. In this paper, we describe a generic approach that improves the information search and retrieval process of different activities with the use of context information. One main goal is the integration of the "crowd" at different stages of this process by combining collective intelligence concepts with context-aware systems. This combination can be used to automatically reduce information overflow by filtering irrelevant data. Furthermore, a real-time information retrieval process without manual search impulses is provided. We also present a prototype as a proof of concept in order to validate feasibility and benefit.**

Keywords - Context Detection; Context Awareness; Information Retrieval; Information Search Process; Collective Intelligence; Crowdsourcing.

I. INTRODUCTION

Nowadays, the importance of information is at a very high level. At the same time, the availability of data (including potentially useful information) is given at any time and location. One major problem is the efficiency of information search and retrieval processes. A study by Delphi Group [1] predicted that employees spent a large amount of time in searching for information (see Figure 1). Only 10% of the respondents disagreed to the statement "Finding the information I need to do my job is difficult and time consuming". The three major impediments finding suitable information are identified as:

- Information changes constantly (41%)
- I don't have good search tools (26%)
- I often don't know exactly what I'm looking for (13%)

Another problem is the fact that a great number of people have suboptimal strategies while using web search engines in order to find relevant information.

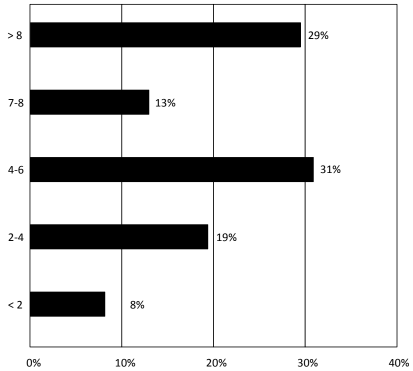


Figure 1. Hours per week spent searching for information [1]

According to a study by Nielsen Norman Group [2], only 1% of the tested persons change their search strategy if the first search results do not fit to the entered issue. One further result of the study is that "users have extraordinarily inadequate research skills when it comes to solving problems on the Web" [2]. Moreover, the study shows that the advanced search features of search engines are not used by most of the test persons "And when they do, they typically use it incorrectly - partly because they use it so rarely that they never really learn how it works" [2].

The main objective of the approach we present in this paper is to handle these impediments by improving the information search and retrieval process using a combination of collective intelligence and context-awareness. Furthermore, the approach addresses passive information retrieval, where people can receive information without even searching for it, respectively receiving solutions for problems users are not aware of. Based on general concepts, the approach can be used in different domains. In Section 4.1, we present three scenarios that show the capability and flexibility of the approach.

The paper is structured as follows. Section 2 introduces current concepts according to context aware systems and our adaption and implementations. In Section 3, we discuss the usage of collective intelligence in order to optimize the information search and retrieval process. Section 4 presents different scenarios and a platform that is used as proof of concept and as a base for the evaluation and validation of the developed approach. Finally, Sections 5 and 6 draw concluding remarks and present related and future work in this area.

ISBN: 978-1-61208-202-8

II. Context Detection and Usage

Several definitions of the word “context” can be found in literature. Some of them primarily refer to location, environment, identity, time or situation [3][4]. An often referenced and more detailed definition is given by Abowd et al.: “Context is any information that can be used to characterize 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 applications themselves [5].”

Concepts, where concrete context information is used to influence a behavior of systems, devices or environments, have been in the focus of research for several years. Particularly in the field of pervasive or ubiquitous systems context information like location or time is often used to change environment behavior. The implementation of context-aware systems depends on the concrete domain. A suitable approach for architectures of context-aware systems in distributed systems is presented by Chen as the Middleware infrastructure [6]. According to this approach, Baldauf et al. identify a common architecture and present a layered conceptual framework for context-aware systems (see Figure 2) [7].

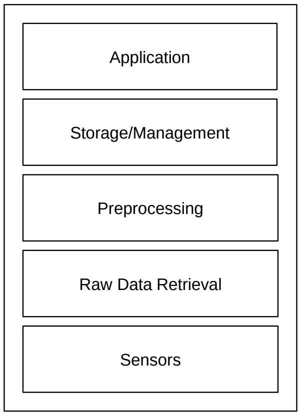


Figure 2. Layered conceptual framework for context-aware systems [7]

At the first layer, sensors collect usable information concerning the concrete context. These sensors can be physical, logical or virtual [8]. The collected data from all sensors (raw data) is transferred to the second layer. The Preprocessing layer includes functionality to convert the raw data to useful information (e.g., reverse geocoding from GPS coordinates). The Storage/Management layer offers the gathered data to the Application layer in an organized form. Finally, the Application layer itself represents the client that uses the detected context information for application functionality. Based on this layered conceptual framework, we developed our novel collective intelligence driven approach.

A. Sensors

As mentioned above context-aware systems use sensors in order to detect useful information inside a concrete environment. In [9], we identified three different environments in the field of software applications:

- *Black-Box Environments* allow no access to the underlying structure (e.g., applications with no access to the application logic or source code)
- *White-Box Environments* allow full access to the underlying structure (e.g., source code)
- *Gray-Box Environments* offer limited access (e.g., support for plug-ins)

The different environments require different sensors for context detection. We differentiate between global sensors and local sensors. Global sensors are high level sensors, e.g., integrated in an operating system to provide global sensor data. Local sensors are low level sensors, e.g., integrated into a concrete application providing application specific sensor data. The data, measured by the sensors, is used in a rule specification process that is described in the following chapter.

B. Rule System

Rules connect sensor data with functionality respectively information, and thus allow mappings between information and concrete user contexts. Considering the crowdsourcing aspect (see Section 3), we created a rule system that is easy to use, but at the same time powerful to support sophisticated context mappings. We use the boolean algebra syntax (with the operations AND, OR, NOT) in combination with fuzzy logic concepts. This provides the creation of simple rules as well as complex rule definitions that are at the same time human readable (see Figure 3).

The structure shows that a rule is a combination of (sub-) rules and sensors. Sensors measure different environment situations. The definition of a sensor includes id, version, value, data type and valid operators. Fuzzy logic information can also be integrated in the sensor definition. Figure 4 shows an XML-representation of the rule presented in Figure 3. One of the sensors (id 78) includes linguistic fuzzy logic expressions in order to compare values to intervals (operator=”equals” value =”high“).

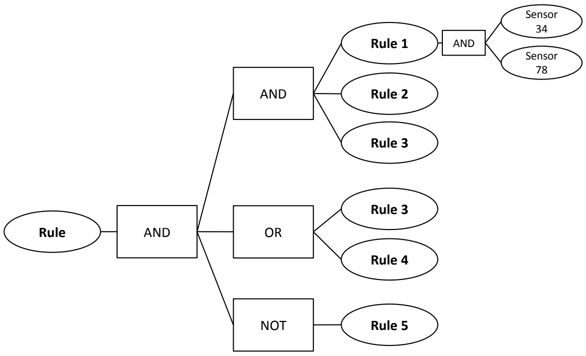


Figure 3. Rule Structure

```
<Rule id="199">
  <AND>
    <AND>
      <Rule id="1">
        <AND>
          <Sensor id="34"
             operator="equals" value="high" />
          <Sensor id="78"
             operator="greaterthan" value="50" />
        </AND>
      <Rule id="2" />
      <Rule id="3" />
    </AND
    <OR>
      <Rule id="3" />
      <Rule id="4" />
    </OR>
    <NOT>
      <Rule id="5" />
    </NOT>
</AND>
</Rule>
```

Figure 4. XML representation of rules

The rule system is decoupled from the sensor system. This allows a separation of sensor creation and rule creation. Sensor creators develop sensors and publish sensor information. Rule creators only need to know information about available sensors and associated sensor data in order to create their rules. Thus the specific know how of the sensor creators (e.g., sensor programming and application integration) as well as the strengths of the rule creators (e.g., expert and process knowledge) can be used in the best way.

C. *Context Mapping*

The relevance of identified information depends to a certain extent on the mapping between sensor data and information. At current status the presented approach provides the mapping types *Explicit Mapping*, *Implicit Mapping* and *Search Query Generation*.

The three mapping types provide both directions of information retrieval, active and passive, respectively information pull and information push. Explicit Mappings support mapping to concrete information, e.g., a document path or a web site. This mapping type provides the best quality of relevance (highest rating). Implicit Mappings map sensor data to application functionality. In a separate process information (e.g., related documents, videos or forum messages) is mapped to application functionality. This concept allows an indirect mapping of context and information. Authors can map their created documents to concrete application functionality. If a rule exists which includes a mapping of sensor data to this functionality the related documents will be displayed.

The third mapping type uses search query generation in order to detect relevant documents. According to the user studies described in [2], we use the collected context information to automatically create queries that use the benefits of a concrete search engine, e.g., advanced search parameters, multilanguage search, use of synonyms, etc. Therefore, in many cases the generated queries are more purposeful than user created queries.

D. *Structure*

Figure 5 shows an extract of structure with relations between context relevant elements. Users act in environments and can own a specific role in this context. They have a need for information that can either be conscious or unconscious. Environments offer a range of functionalities and are observed by sensors. Rules include functionalities, information and sensors. The crowd plays a major role in this structure. It can influence nearly all important elements of the approach. People, that are part of the crowd, develop sensors, define rules, specify functionalities and create information of different types. Advantage of this structure is that the different activities are independent from each other; the particular skills of the crowd members can therefore be used ideally. In the next chapter, we present concepts for the integration of the crowd into the activities.

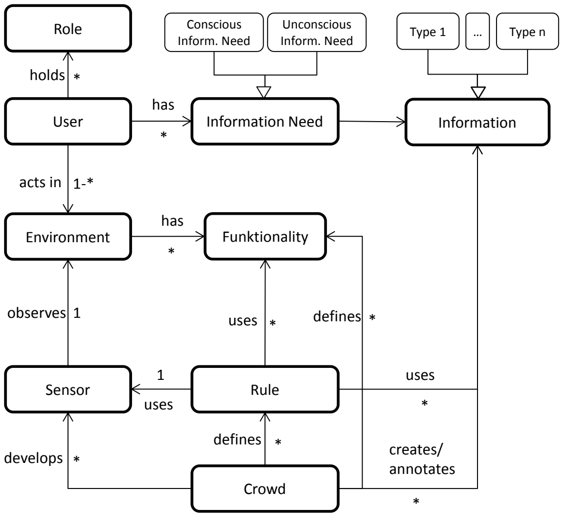


Figure 5. Main Elements and Relations

III. CROWDSOURCING

A major goal of the presented approach is the integration and participation of the mass (crowd). This affects the information retrieval process as well as the context detection and metadata enhancement. Research activities, concerning the "potential of groups", have been in focus for a couple of years and in different domains, e.g., biology, social science and computer science. The generic term of this research discipline is *Collective Intelligence*.

A. *Collective Intelligence*

Levy defines Collective Intelligence as a „form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills. [...] The basis and goal of collaborative intelligence is the mutual recognition and enrichment of individuals rather than the cult of fetishized or hypostatized communities [10]." Because of the recent internet infrastructure, collective intelligence has gained in importance in different scenarios. "Collective Intelligence has received a new meaning in recent years, especially through the emergence of new (mostly Web 2.0) applications and user generated content [11].”

Malone et al. identified a small set of building blocks according to most types of collective intelligence systems. In order to classify these building blocks four key questions have to be answered (see Figure 6) [12]: *Who is performing the task? Why are they doing it? What is being accomplished? How is it being done?*

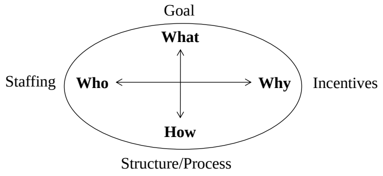


Figure 6. Elements of collective intelligence building blocks or “genes” [12]

According to the approach we present in this paper, the answers to the key questions are as follows:

Who is performing the task? – The crowd, which is represented by an independent mass of people [13]. Participating persons can hold different roles, e.g., author of a document, expert inside a forum/domain, rule creator or information/functionality mapper. They all are part of the crowd and can collectively optimize the entire process.

Why are they doing it? – In most recent collective intelligence systems, the motivation of participation is founded in “Money”, “Glory” or “Love” [12]. In the presented approach another major reason to participate is the own benefit of the results. On the one hand, this applies to the quality and relevance of the received information (reader’s view). On the other hand, the participation provides a targeted distribution of relevant documents (author’s view).

What is being accomplished? – The crowd can affect the quantity, quality and relevance of the detected information in different ways and with the accomplishment of different activities. With regard to the presented approach, main activities concern sensor development, functionality and rule definitions as well as metadata annotations. The next sections describe the main activities in detail.

How is it being done? – We developed a platform that integrates different activities and artifacts of the presented approach. Sensors can be published, rules can be defined and mappings can be provided. The data is stored in a shared database. The local tool for user interaction and information presentation also uses this database (see Section 4) in order to provide real-time changes.

B. Creation of Sensors

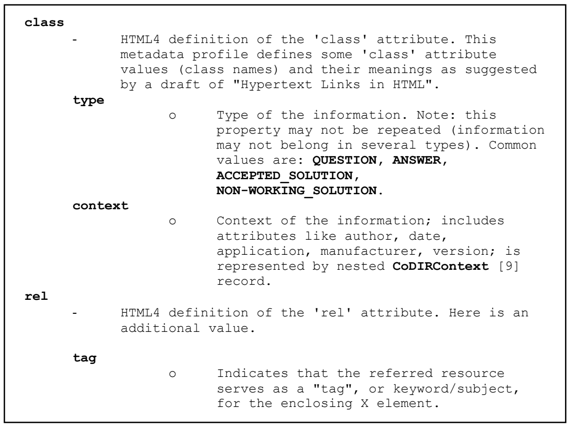
In order to analyze a concrete situation of users, sensors are required that measure data inside environments. Our approach provides the creation of (virtual) sensors by the crowd. Therefore, we defined a one way interface between sensor and processing component. The communication is realized as a simple REST-request containing key value pairs of sensor data. This concept provides an open and technology independent sensor development. According to Black-Box environments (see Section 2.1), tools can be developed that are able to catch user interaction without intervention to the application itself. Furthermore, the crowd can develop and share plug-ins for using in applications of type Gray-Box. The most efficient way is sending sensor data directly from the application logic. Our approach supports the integration into the application architecture with minimal intervention.

C. Context Mapping

As described above, a context describes a concrete situation of objects. Strang and Linnhoff-Popien identified six types of modeling context which cover different requirements: *Key-Value Models, Markup Scheme Models, Graphical Models, Object Oriented Models, Logic Based Models* and *Ontology Based Models*. While the Key-Value Models cover least requirements, the Ontology Based Models support most of them [14]. We use the Key-Value-Model to collect required data of the sensors. In order to map functionality to aggregated sensor data, we implement ontology-based models and adapted the approach from Wang et al. [15] for the creation of concrete context information for a wide range of domains and scenarios. Using a high-level ontology, the crowd can create ontologies for a concrete domain (e.g., application). The ontologies are also stored at the shared database and can be accessed by the crowd in order to share and optimize it.

D. Enhancement of the Information Source

An additional way to optimize the search result quality is the enhancement of documents with metadata. Nowadays, relevant documents are often web-based and belong to categories like websites, blogs, expert systems, wikis, forums, tweets or social communities. In order to integrate the crowd into the annotation process of relevant artifacts, an efficient practice has to be available; otherwise the willingness to participate decreases. Metadata annotation concepts like RDFa [16], microformats [17], Microdata [18] and Schema.org [19] cover these requirements. Currently we are developing a schema for forums and expert systems where the different types of entries can be annotated with additional information. These types are, e.g., questions, answers, accepted solutions and non-working solutions.



```
class
      -     HTML4 definition of the 'class' attribute. This
            metadata profile defines some 'class' attribute
            values (class names) and their meanings as suggested
            by a draft of "Hypertext Links in HTML".
      type
                  o     Type of the information. Note: this
                        property may not be repeated (information
                        may not belong in several types). Common
                        values are: QUESTION, ANSWER,
                        ACCEPTED_SOLUTION,
                        NON-WORKING_SOLUTION.
      context
                  o     Context of the information; includes
                        attributes like author, date,
                        application, manufacturer, version; is
                        represented by nested CoDIRContext [9]
                        record.
rel
      -     HTML4 definition of the 'rel' attribute. Here is an
            additional value.

      tag
                  o     Indicates that the referred resource
                        serves as a "tag", or keyword/subject,
                        for the enclosing X element.
```

Figure 7. Microformat profile (draft) for forums and expert systems

The additional metadata allows machines to filter non-relevant information according to a specific context. Figure 7 shows a proposal of a microformat profile for forum-based expert systems. This profile uses the CoDIR-microformat [9] as nested attribute.

IV. PROOF OF CONCEPT

The following scenario demonstrates the usage and implementation of the presented approach in the context of software engineering activities inside the disciplines SE-Development and SE-Tools.

In order to validate feasibility and benefit of the presented approach, we developed and enhanced a prototype as a proof of concept and added support for tasks in software engineering activities. Figure 9 illustrates the core elements of the prototype.

Virtual Sensors collect context relevant information inside a specific environment. If a sensor detects a change of context, the new context is transferred to the Result & Interaction Interface (RII). The interface transfers the context data and additional parameters to the Information Broker. The Information Broker uses the context information, different search providers and repositories in order to search for relevant information. The results (received from the search providers and enhanced with ranking information) are send back to the RII and displayed to the user. The user can directly access the received resources and (optional) rate the relevance of the information in relation to the current context.

In the presented real-world scenario, the crowd has published a plugin for the IDE Eclipse to track user activities and system behavior. This plugin represents a virtual sensor. This virtual sensor is able to collect different information about the current context based on a domain specific ontology published by the crowd.

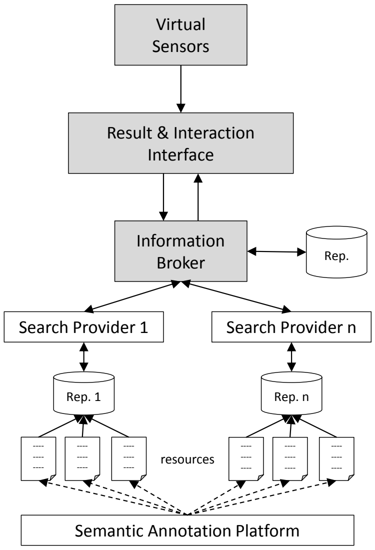
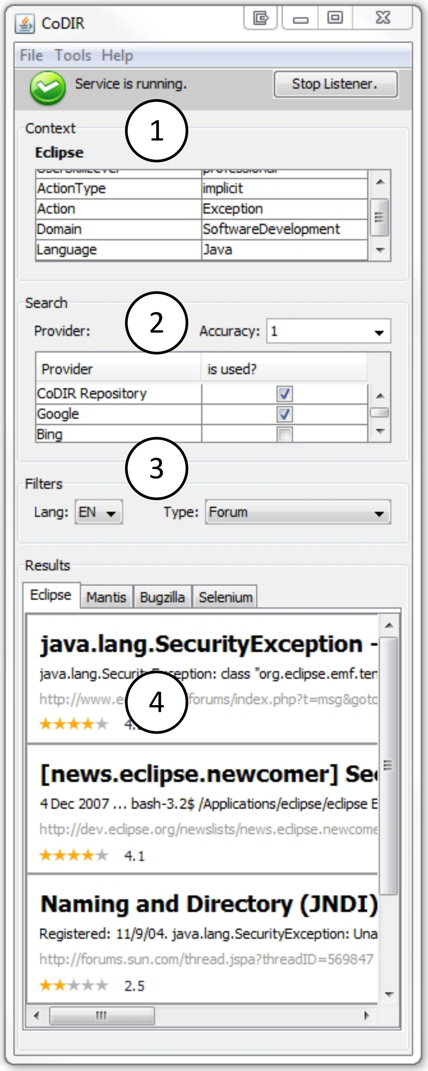


Figure 8. Core elements of the prototype

If the application for example throws a security exception, a context-object is generated and directly transferred to the RII. The RII is implemented as Rich Client Application (see Figure 8). At the top the details of the current context according to the context ontology model are displayed (1). The user can influence the search behavior by customizing different parameters, like the available search provider (2), the language or the document types (3).

The RII transfers all useful information to the Information Broker, which is implemented as a web service. We also implemented a Web Portal for publishing sensors, rules and functionality/information mappings. The Information Broker uses the context information and parameters to search for relevant resources, and transfers the results back to the IRR (4). Inside the result list, the user can rate the relevance of the result items in order to optimize the future search activities for all users.



Service is running.

Context

Eclipse

ActionType	implicit
Action	Exception
Domain	SoftwareDevelopment
Language	Java

Search

Provider: Accuracy: 1

Provider	is used?
CoDIR Repository	☑
Google	☑
Bing	☐

Filters

Lang: EN Type: Forum

Results

java.lang.SecurityException -
java.lang.Securi ception: class "org.eclipse.emf.ter
http://www.e forums/index.php?t=msg&gotc
★★★★☆ 4.

[news.eclipse.newcomer] Se
4 Dec 2007 ... bash-3.2$ /Applications/eclipse/eclipse E
http://dev.eclipse.org/newslists/news.eclipse.newcome
★★★★☆ 4.1

Naming and Directory (JNDI)
Registered: 11/9/04. java.lang.SecurityException: Una
http://forums.sun.com/thread.jspa?threadID=569847
★★☆☆☆ 2.5

Figure 9. Screenshot of the Result & Interaction Interface (RII)

V. Related Work

Related work can be found in the area of context-aware, ubiquitous and pervasive systems. Several context ontologies have been proposed, mainly with focus on pervasive and mobile computing using physical sensors, e.g., CONON [15], SOUPA [20] and CoBrA Ontology [6]. Our approach addresses domain independent software applications in combination of virtual sensors that are integrated in different types of environments (application systems). According to search activities for relevant documents, several information retrieval approaches have been proposed, e.g., POLAR, a probabilistic object-oriented logical framework for annotation-based information retrieval [21]. Related work in the field of collective intelligence can be found in crowdsourcing approaches [13], Social Web Applications [11] and Web 2.0 technologies. We use the advantages of the different concepts to allow an influence of the crowd in nearly all phases of the process. In [22], Soylu and Causmaecker present an approach of empowering context-aware pervasive computing environments with embedded semantics. Differences to our approach are the missing reference to information retrieval concepts as well as the focused domain.

VI. Conclusion and Future Work

In this paper, we presented a novel approach that enables the improvement of information search and retrieval processes in software application environments. The focus concentrates on software applications, where the demand on real-time information retrieval is given in different disciplines. The participation of the crowd is a fundamental component of our approach. Hence, an important task for future work is to constantly simplify the process for participation in order to enhance the willingness of the crowd. Using the prototype, we identified the need for a suitable visualization (e.g., graphs or maps) of the results as well as customized recommendation techniques. Currently, we are working on a hybrid and multidimensional recommender system that allows transparency and enhanced filter options. Another goal is the integration of our approach into environments that need security and trust features. Users can then be informed if they are in an unsecure context, or if they share confidential documents.

References

[1] Delphi Group (n.d.): Information Intelligence: Content Classification and the Enterprise Taxonomy Practice. Boston, MA (2004)

[2] Nielsen, J.: Incompetent Research Skills Curb Users' Problem Solving. In: Jakob Nielsen’s Alertbox (2011)

[3] Ryan, N., Pascoe, J., and Morse, D.: Enhanced reality fieldwork: The context-aware archaeological assistant. In: Computer Applications in Archaeology. Edited by V. Gaffney, M. van Leusen and S. Exxon. (1997).

[4] Hull, R., Neaves, P., and Bedford-Roberts, J.: Towards situated computing. In: Proceedings of the First International Symposium on Wearable Computers, pp. 146-153, Cambridge, MA , USA (1997)

[5] Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., and Steggles, P.: Towards a Better Understanding of Context and Context-Awareness. In: 1st international symposium on Handheld and Ubiquitous Computing, pp. 304-307. Springer-Verlag. Karlsruhe, Germany (1999)

[6] Chen, H.: An Intelligent Broker Architecture for Pervasive Context-Aware Systems. PhD thesis, University of Maryland, Baltimore County (2004)

[7] Baldauf, M., Dustdar, S., and Rosenberg, F.: A Survey on Context-Aware Systems. In: Inter-national Journal of Ad Hoc and Ubiquitous Computing, vol. 2, nr. 4, pp. 263-277. Inderscience Publishers. Geneva, Switzerland (2007)

[8] Indulska, J. and Sutton, P.: Location management in pervasive systems. In Proceedings of the Australasian Information Security Workshop (CRPITS 03), pp. 143-151, Australian Computer Society. Sydney, Australia (2003)

[9] Beul, M. and Eicker, S.: Don’t call us, we call you. A community driven approach for (domain independent) context driven information retrieval (CoDIR). In Proceedings of Fifth Inter-national Conference on Signal Image Technology and Internet Based Systems (SITIS), pp. 458-464, IEEE Press, New York (2009)

[10] Levy, P: Collective Intelligence: Mankind’s Emerging World in Cyberspace. Translated by R. Bononno. Perseus Books. Cambridge, MA (1997)

[11] Leimeister, J.M.: Collective Intelligence. In: Business & Information Systems Engineering, vol. 2, nr. 4, pp. 245-248 (2010)

[12] Malone T., Laubacher R., and Dellarocas, C.: Harnessing Crowds: Mapping the Genome of Collective Intelligence, Working Paper No. 2009-001. MIT Center for Collective Intelligence, Cambridge, MA (2009)

[13] Howe, J.: Crowdsourcing - How the power of the crowd is driving the future of business. RH Business Books, London, England (2009)

[14] Strang, T., and Linnhoff-Popien, C.: A Context Modeling Survey. In: Workshop on Advanced Context Modelling, Reasoning and Management. UbiComp, Nottingham, England (2004)

[15] Wang, H.H., Zhang, D.Q., Gu, T., and Gung, H.K.: Ontology Based Context Modeling and Reasoning using OWL. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops. Pp. 18-22, IEEE Press, New York (2004)

[16] Adida, B. and Birbeck, M.: RDFa Primer. W3C, http://www.w3.org/TR/xhtml-rdfa-primer, [retrieved: April, 2012]

[17] Microformats.org (n.d.): About microformats, http://microformats.org/about, [retrieved: April, 2012]

[18] Hickson, I.: HTML Microdata. W3C, http://dev.w3.org/html5/md/Overview.html, [retrieved: April, 2012]

[19] Schema.org (n.d.): Getting started with schema.org, http://www.schema.org/docs/gs.html, [retrieved: April, 2012]

[20] Chen, H., Perich, F., Finin, T.W., and Joshi, A.: SOUPA: standard ontology for ubiquitous and pervasive applications. In: 1st Int. Conf. on Mobile and Ubiquitous Systems: Networking and Services, pp. 258-267, Boston, MA (2004)

[21] Frommholz, I. and Fuhr, N.: Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In: Opening Information Horizons - Proc. of the 6th ACM/IEEE Joint Conference on Digtial Libraries (JCDL), pp. 55-64, ACM. New York, NY, USA (2006)

[22] Soylu, A. and Causmaecker, P.D.: Embedded Semantics Empowering Context-Aware Pervasive Computing Environments. In: Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 310-317. Brisbane, QLD (2009)