

Automatic ontology-based User Profile Learning from heterogeneous Web Resources in a Big Data Context

Anett Hoppe

Supervised by C. Nicolle and A. Roxin
CheckSem Group, LE2I
Université de Bourgogne
Dijon, France

[first name].[last name]@checksem.fr

ABSTRACT

The Web has developed to the biggest source of information and entertainment in the world. By its size, its adaptability and flexibility, it challenged our current paradigms on information sharing in several areas. By offering everybody the opportunity to release own contents in a fast and cheap way, the Web already led to a revolution of the traditional publishing world and just now, it commences to change the perspective on advertisements. With the possibility to adapt the contents displayed on a page dynamically based on the viewer's context, campaigns launched to target rough customer groups will become an element of the past. However, this new ecosystem, that relates advertisements with the user, heavily relies on the quality of the underlying user profile. This profile has to be able to model any combination of user characteristics, the relations between its composing elements and the uncertainty that stems from the automated processing of real-world data. The work at hand describes the beginnings of a PhD project that aims to tackle those issues using a combination of data analysis, ontology engineering and processing of big data resources provided by an industrial partner. The final goal is to automatically construct and populate a profile ontology for each user identified by the system. This allows to associate these users to high-value audience segments in order to drive digital marketing.

1. INTRODUCTION

The development of the Web since the 1990s has largely altered the way we perceive, deploy, and exchange information nowadays. The leap in technological process had the power to change whole industries, in most cases giving more control to the customer, who, on the web, is able to individually choose his/her contents of interest and interact directly with the information.

The area of interest tackled in the presented project is the advertisement domain. Advertisers have always been

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Articles from this volume were invited to present their results at The 39th International Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy.

Proceedings of the VLDB Endowment, Vol. 6, No. 12
Copyright 2013 VLDB Endowment 2150-8097/13/10... \$ 10.00.

keen to identify those subjects from a bigger audience that are open for the message of a certain publicity – and to lead them to the facilities of the respective business partner. The technology of the Web offers a whole new range of possibilities to examine and qualify customers individually. Whereas, for instance, print ads enabled generic targeting the audience of a certain newspaper, online advertisements can be placed for each individual, based on the available knowledge about him/her.

The PhD project presented on the following pages is realised hand in hand with an enterprise that provides targeting-relevant data analysis to mayor online publishers in France. It will explore an ontology-based approach for user profiling that integrates (a) the knowledge existing within the company in an explicit modelisation, (b) new insights gained from automatic content analysis. Novelties therein are the adoption of ontologies as underlying data structure through the whole analysis process, and the usage of ontological inference for the attribution of data sets to a certain customer segment. Those steps will be supported by the integration of state-of-the-art content analysis techniques. Their adaptation to the specific problem task includes the adjustment for the usage on French language texts and the customisation for the problem domain.

The following sections give a more precise description of the context of the research (Section 2), along with an overview of relevant related work (Section 3) and finally of the solutions pursued.

2. PROJECT DESCRIPTION

The work to be done in the present PhD is carried out in the realms of a joint project between the Checksem research team at Dijon, France and an IT enterprise that is based in Bordeaux, France.

The Checksem research team is a transversal project within the LE2I research laboratory of Université de Bourgogne, Dijon, France. It comprises about 25 people in different stages of their research career. The shared interest of research is semantic interoperability in heterogeneous information systems. Specifically, team members develop solutions for applications in the domains of recommendation systems, building management and geographic information systems.

The industrial partner currently employs about 30 persons. The company unites experts in statistical analysis, machine learning, natural language processing and big data administration. One of the main goals is to exploit the additional opportunities offered by the Web to introduce a

new generation of web advertisements to the French market. Whereas the current scheme allows advertisers to target a rough consumer group by choosing the web page that their content gets published in, the company employs various analysis techniques to build individual user profiles. This allows to automatically decide if a single user belongs to a certain marketing segment. Thus, advertisers may dynamically attribute the suitable content.

All analysis is based on the user cookies provided by each one of the about 120 publisher pages associated to their clients. The data include the click stream that the user effectuated on the client page, the time spent on each page and typical background information. Based on the sequence of articles viewed, one can deduce main interests expressed along with the logical relations drawn between them.

Given this context, the task at hand comes with three key issues:

1. **Background analysis:** The analysis of the web resources viewed by the user are a crucial step in constructing the user profile. Their content has to be unambiguously analysed in order to be transformed into a machine-interpretable representation.
2. **User profiling:** Based on all available information, a unique profile for each user has to be constructed. The background knowledge about relations between user characteristics has to be extended with knowledge that can be deduced from the user's specific trace.
3. **Big data scale:** Cookies assemble the event traces for the users of each partner site. This leads to a high volume of heterogeneous data that have to be interpreted. Every solution thus has to be efficient and scalable.

The following paragraphs take a closer look at each of the above-identified problem areas to give a more descriptive view on their respective keylocks.

2.1 Background Analysis:

The main source of information will be the textual resources linked to the history of the profiled user. The analysis of text has been widely studied through different domains, such as Information Retrieval and Natural Language Processing. Efforts have been made with focus on the extraction of meaning from text and its transformation to machine-interpretable structure. However, the toolkits available often centre around the evaluation of English resources and offer only limited or no support for other languages. As most of the clients of the partner company publish in French, the adaptation of state-of-the-art approaches to French language will pose one of the mayor keylocks. Furthermore, we will explore extensions of their functionality that serve their customisation for the problem task, focussing on the extraction of information that are significant for market-relevant customer segmentation.

2.2 User Profiling:

The overall goal of all analysis is to derive a valid user profile that will be the base of the audience segmentation. One crucial working step is thus the automatic and correct deduction of user interests from their navigation history.

Traditional approaches for user profiling often rely on sets of keywords for the summary of user preferences [2]. This

representation scores with its simplicity and thus the possibility to be efficiently computed and processed. However, it fails to capture all of the subtleties of natural language, as for example polysemy (the fact that same-written words might adopt different senses, depending on their context of usage), or synonymy (two different words relating to the same semantic concept).

In the MindMinings project, we aim for the use of ontologies and Semantic Web resources as underlying data structure and background knowledge respectively. To the best of our knowledge this approach is novel. Even though there exist recent publications (for more information, please refer to Section 3) that propose ontology-shaped profile representations, all of them fall back to the keyword paradigm at one or another moment in the analysis process, particularly when it comes to distance computations. We use ontologies through the complete profiling process, for the representation of the content analysis, the user interests and the inference of the correct, affiliated audience segment. Furthermore, the ontology serves as an explicit representation of the knowledge of the company-internal context knowledge and as an integration point for their own, machine learning-based data analysis and our semantically-enhanced content information.

2.3 Big Data:

Besides the highest possible correctness, the profiling approach has also to measure up to another demand of the industrial partner: The interpretation of incoming events, the dynamic creation of the user profile and its exploitation for content suggestion have to scale to the number of partner sites and the number of users that surf those pages every day. Given the number of partner sites and the nature of the data, we can in all conscience claim to work in a Big Data-context. According to the IEEE¹ and several research works [10, 7], the following attributes qualify an analysis task to be settled in a Big Data domain:

1. Volume – The amount of data assembled by the company each month reaches about 150 million user events per month for an average partner, in sum 2,4 billion events each month, leading to a high volume of data to analyse.
2. Variety – Our analysis includes web data in all its formats and orientations, leading to a highly variable spectrum.
3. Velocity – User events arrive for analysis in real-time, meaning a high frequency during main activity hours.
4. Value – As the analysis is performed on real-life data, generated by the activity of users, it bears information about their interests and habits that is commercially exploitable.
5. Veracity – Stemming from the above named criteria, a fifth one arises: When data instances arrive in a crazy speed and in various formats, ambiguities and faulty entries have to be expected. Thus, trustworthiness of a data stream and a measurement of uncertainty are points to consider within the analysis.

¹<http://www.ieee.org>

The profiling task at hand qualifies in all criteria. User clicks appear in a high number and also in high frequency. Starting from the numbers indicated above, on an average day, all current clients would produce a number of 80 million data instances scattered over day periods with different degrees of activity. Even though data arrives in form of uniform user events, the web resources referenced therein expose a high level of variety and may include different types of media (articles, blog entries, pictures, videos, etc.) and basically all combinations of topics. The value of the information included, however, is a better understanding of the user behaviour and deduce interests. However, in such a stream, erroneous events will certainly happen. This refers to all click data that are reported, but that are not pertinent for the user's profile.

2.4 Uncertainty Handling:

The analysis of natural user behaviour implies an important challenge: human beings seldom act upon logical, crisp principles with clear boundaries. Interests and preferences change, even on hourly basis, depending on life situation, habits or moods. As a result, every pattern observed in a user's behaviour will have to be carefully examined for its meaning and importance, before including it into the final profile. Supposing a user opens, for the first time in his/her navigation history, an article about soccer – this might indeed signify a newly developed interest; it might, however, also signify that another person is using computer for a few minutes, or an erroneous link tag, or a misleading headline. A framework is needed that allows to include uncertainty information for each observed fact, and that allows inference on those uncertain statements – leading to conclusions that may or not be true based on the amount of evidence leading to them.

One theoretical framework to display those notions of uncertainty is fuzzy logic. First introduced by L.A. Zadeh in 1965 [22], it includes theoretical background to quantify the vague aspects of an assertion and assign resulting uncertainty value to all inferred statements. The inclusion of fuzzy logic to ontological structures has a rather short history, with a first comprehensive work having appeared in 2006 [16]. However, first computing schemes and applications have been presented therein. In the context of this PhD, fuzzy techniques may be an answer to the question how to include uncertainty to the ontological profile representation.

2.5 Summary

The task of this PhD project thus classifies as a combination of several research domains. The scope comprises relations to the areas of machine learning, data mining, and natural language processing. The information that are relevant for the profiling process have to be discovered from mainly textual resources and integrated to a pre-defined data structure that is not only to capture the results of this content analysis, but also the results stemming from the company's internal analysis and facts that can be directly parsed from the navigation history.

3. RELATED WORK

3.1 Website Qualification

As already suggested in Section 2, there has been a considerable amount of work on the qualification of web resources, even including Semantic Web resources as a matter of reference.

In [13], the inclusion of a Semantic Web reference resource has proven to boost efficiency. The authors compare the traditional statistical LSI approach (“Latent Semantic Indexing“ [6]) for topic discovery with a WordNet-based alternative. They proved the latter to achieve results as good as the traditional approach, but in considerably lower computation time.

A large number of works report the success of the inclusion of WordNet [11], DBPedia [1] and specific domain ontologies to the pre-processing of textual resources. [14, 15] review techniques for Named Entity Recognition and name structured knowledge sources as a supportive element for the task, besides unstructured sources as corpora. [5] uses a combination of a semantic topic attributed to a term and his syntactic role within the sentence to boost performance for Part-of-Speech tagging.

Furthermore, Semantic Web resources have been applied for both, feature reduction and query extension. Both techniques tackle the problem of feature sparseness – given the high number of words in natural language, one text might only feature a small number of the words that are contained in the dictionary of all documents. Besides, the existence of synonyms in natural language might lead to texts that treat the same topic, but show no overlaps in their used vocabulary. In consequence, a regular matching algorithm would not find them to be similar. Anyhow, the background knowledge about semantic relations that is contained in a semantic dictionary as WordNet [11], can be used to either (a) cluster words that are closely related and thus reduce the feature space, or (b) extend each keyword vector by the synonyms of the terms already included and thus facilitate a matching between documents that use disjoint, but synonymic sets of words.

However, approaches concentrate around improvements of the traditional keyword based approach. The quality of topic terms benefits from disambiguation or extension by synonymic terms. The semantic relations of the concepts are thus used to ameliorate a traditional paradigm, but then omitted for further exploitation.

3.2 User Profiling using Ontologies

As already described above, the traditional methods (keyword-based profile representations) and their extensions (e.g. semantic networks) do not suffice to capture all subtleties caused by the properties of natural languages (e.g.: polysemy, synonymy). Presented approaches use background knowledge in form of taxonomies and/or ontologies to bridge this gap. Notably this is done by identifying the correct sense of a term with respect to its context and relating it to a Semantic Web resource that provides a close definition of the specified meaning. This association is done by means of uniform resource identifiers (URIs) or specific predicates from ontology languages (e.g.: `owl:equivalentClass`²). User interests are therein mapped to the concepts of the domain ontology and thus explicitly and unambiguously identified.

However, recent approaches go a step further by including the semantic relations that are identified between the concepts of the ontology into the profile – and thus, extending

²<http://www.w3.org/TR/owl-ref/#equivalentClass-def>

it from a set of keywords to a *profile ontology*. This may be done using further background knowledge, or just using the sequence of user actions to interconnect topic terms [2]. The first type of methods largely relies on data mining and fuzzy relational algebra – the co-occurrence pattern of terms is used to predict a relation between them for the current user’s context [23, 9]. In the second group, often used background resources are the Yahoo Web directory³[8] and the ODP⁴ (Open Directory Project)[17]. Whereas the Yahoo directory is limited to hierarchic relations, thus providing a taxonomy of terms, the ODP classification is often referred to as an ontology [2], [12] criticises the extension of the taxonomic relations by “symbolic”- and “related”-relations as too shallow and thus suggests that the ODP can at most be considered a light ontology. [3, 4] suggest composing graph-based profiles for all queries that are related to one search session. The initial query is mapped to the concepts of an ontology (ODP in this case) and used as a seed profile, that subsequently gets extended by the vocabulary attributed to each new arriving query. [18] analyses the web pages viewed by a user on the base of domain ontologies to build a personal profile. However, they demand the user to explicitly ascertain each newly added term to exclude noise from the profile ontology. In a recent work, [2] proposes to use the semantic information included in WordNet to map topic terms to concepts of the Yago ontology [20]. [9, 21] feature a combination of implicit and explicit user information – first asking the user for topic concepts that provide the skeleton for the personal ontology, and then extending the notions by mined information that are classified using the standard of the LCSH (Library of Congress Subjects Headings).

In summary, we can state that besides the very recent work of Calegari and Pasi [2], all representations tend to use taxonomic information or just light ontologies as the ODP. However, yet those approaches show promising results that may be further improved by employing a rich domain ontology as the base for analysis.

4. OUTLINE OF PURSUED SOLUTIONS

As already evoked, the main goal of the described PhD thesis is to present novel solutions for user profiling based on web navigation history. In a first stage, a prototype has been developed using start-of-the-art techniques and tools from the concerned domains. In the course of the project, this prototype will be the base for extensions and adaptations and their evaluation.

4.1 Online and offline tasks

The in-depth analysis of textual resources can be a costly task that, if performed on-the-fly, might endanger the performance of our system. The amount of data that has to be processed with a request of the industrial partner that demands real-time capabilities of the profiling system led to an adapted design decision. Even though the partners of the industrial partner are numerous, they still own a limited and defined set of websites. Therefore, we decouple the web page qualification from the actual profiling task to come up to real-time performance (an approach similar to the indexing step that enables Information Retrieval systems to answer user requests in short time spans). As soon as a new client

³<http://yahoo.com>

⁴<http://www.dmoz.com>

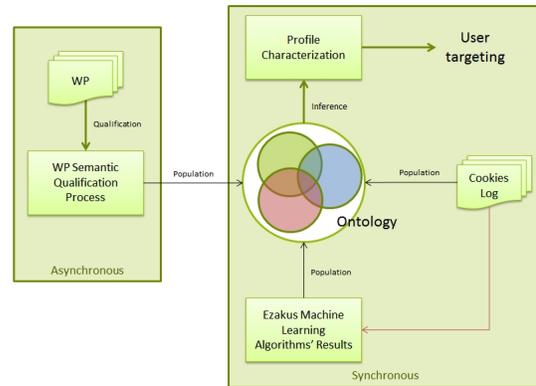


Figure 1: Overview of the information that will be used to populate the ontology (separated offline and online processes)

appears, his domains are registered, analysed and the resulting information stored in the ontology. These information are, of course, continuously updated to capture the evolution of the site’s content. During the evaluation of the user navigation information these stored information are retrieved and connected based on their sequence and additional context information. The latter complement the internally conceived semantic information from the web page analysis. It stems from two different sources: (a) facts derived directly from the navigation logs, such as the user agent, (b) results from the company’s internal analyses. Figure 4.1 shows a generic overview of the information sources contributing to the ontology population, including the separation of offline and online processes.

The following passages review in detail the techniques employed for the page analysis, followed by a reprise of the issues arising when integrating the subsequent information to the ontology and with their updating over time.

4.2 Basis: The ontology

In close collaboration with the industrial partner, we developed an ontology that models the informational context that it is working in. It does not only capture the entities directly touched by the analysis process, such as the web page, the keywords and the respective topic categories, but also concepts that stem from the commercial context of the enterprise, such as the clients and their domains. Additionally, we included an entity *Weight* that enables the modelling of membership degrees for non-binary relations. This is applied, for instance, when modelling the relation between an article and its topic categories. We are thus able to include the information about the category affiliation of an article, but also, how close the relationship among them is. Practically, each non-binary relation will be modelled as a concatenation of two sub-relations: one leading from the initial entity (e.g. the article) to the *Weight* entity, and one leading from there to the destination entity (e.g. the topic category).

4.3 Offline: Content Analysis

For a proof of concept, we designed a first prototype of the

information extraction module that relies on well-established techniques. The development serves as a validation of the workflow and as a reliable baseline solution to evaluate our extensions.

Therein, each web domain that belongs to a client is registered – a set that is at the moment limited a set for testing purposes. Using the links between the pages, the underlying web pages are discovered and undergo an initial examination to determine their value for semantic evaluation. This filtering step allows to optimise performance – pages that do not contain sufficient semantic content will be excluded from the subsequent, more costly analysis steps. For example, this step excludes the overview pages that are often employed by news sites. These pages feature headlines and snippets from several domains to offer the reader an entry point to all contents. But requesting them does not reveal much information about a user. In contrast, much more information is contained in the leaf pages that feature the actual articles with all their semantic meaning and its evidence about a user’s interests. After this filtering step, the profound analysis of the left pages is initiated. A parser retrieves the textual components. So far, keywords are extracted using the traditional tf-idf measure. The resulting set of terms is complemented by named entities identified by the tool OpenCalais⁵. The sum of both is used to relate each text with a content category.

The categories build the basis for the identification of a user’s interests. They have been defined in cooperation with the company to reflect the relevant topics exposed in their partner sites. Each of them is attributed with a set of keywords as features. A comparison between those and the keywords found in the articles enables the computation of a similarity value that is the base for an affiliation of a certain category to an online article.

These developments combine current, but well-established techniques. Improvements will be achieved by introducing semantics to the analysis process and by adapting the algorithms to (a) French language, and (b) the problem domain. Whereas there are well-established technologies and tool kits for key term determination and named entity recognition, most of them focus on the analysis of English text. For example, the above-cited framework for Named Entity Recognition, OpenCalais, offers to date the discovery of 15 different types of named entities for French language, in comparison to more than 120 different entities and events for English. Furthermore, all tools offer generic discovery of entities. A future extension could be the exploration of the possibility of a specialised key term extraction for advertisement-relevant terms – based on the domain knowledge of the industrial partner.

For the distance computation between an article and the topic categories, a standard vector-based approach has been employed. We aim for a semantic distance measure, as soon as the discovery of the correct semantic counterpart for each ambiguous keyword has been successfully integrated to the information extraction step. Tests with a distance measure based on the DBPedia graph showed promising results already for the comparison of simple, non-disambiguated keywords. The performance of the system should further increase when entering correct links to the entity that is actually referenced in the text.

⁵www.opencalais.com

4.4 Online: Information integration and inference

All information that is available prior to the profiling will be stored during the offline process. However, the processing of the user information will have to be done online, mainly the integration of the navigation log information and the deduction of the user profile.

The integration step includes user information (a) that can be extracted directly from the cookie, such as the user agent and (b) the results of the company’s internal analysis operations. The latter are obtained using enhanced machine learning techniques and statistical analysis based on the user navigation. Their integration will be realised using an agreed-upon exchange protocol.

Those information at hand, the inference of the correct segment affiliation is the big challenge for the system. On the base of the enterprise’s know-how about valid customer segments and their attributes, OWL inference and SWRL rules will be used to deduct the correct segments for each user. Therein, the inclusion of the proportional membership using the weight concept will force us to rethink the evaluation of the concerned relations. From the mathematical frameworks that enable the management of notions of vagueness, fuzzy inference is a promising candidate for our application. However, this design decision has not yet been fully implemented and tested.

Another aspect that will raise interesting research questions is the integration of the temporal facet to our content analysis. The user’s browsing behaviour may reflect long-term interests and short-term information needs that have to be integrated to the ontology as such. Thus, based on the time stamps, we want to identify those types of information and shape the weighting of the profile contents accordingly. Similar research projects the Spreading Activation Theory for this purpose [2, 19] and hence, offer us a starting point for our own considerations.

5. SUMMARY/CONCLUSION

The article at hand presents the work envisioned for a PhD project, to be realised between October 2012 and October 2015 at the University of Burgundy, Dijon, France. All research done is supported by two collaboration partners, the Checksem working group, as a part of the LE2I laboratory at Dijon, and an IT enterprise based in Bordeaux, France. The intent is to develop a novel approach for user profiling, based on the information from user surfing logs. Those logs are obtained from the commercial clients of company and currently comprise 120 web sites, mainly in French language. Their analysis is carried out respecting the users’ privacy and treating each client’s information confidential.

The novelty of the approach once consists in the consequent use of ontologies for knowledge representation, in all stages of the analysis process. Thus, both, the web page profile and the user profile are represented in form of explicit and formally defined linked concepts. Potential ambiguities will be resolved by in-depth pre-processing and by relating terms in doubt with concepts defined in freely available semantic datasets. Thereby, the processes of content evaluation and profile construction are effectuated without user involvement, only relying on the implicit information that is contained in each user’s browsing log. All analysis steps

have to be scalable to usage in a big data context and will thus be thoroughly tested for their efficiency.

In its application, the user profile will serve a better understanding of user's characteristics. As the Web offers the possibilities to analyse each user's behaviour individually, a thorough modelisation of his/her context and preferences enables dynamic composition of suitable content. In the short run, this facilitates a site owner to gain precise knowledge about the audience reached with a certain content. In a near future, this information may be used to display the most relevant adverts for a given profile and thus driving digital marketing.

6. ACKNOWLEDGEMENTS

This research is supported by Ezakus Labs HQ, www.ezakus.com.

7. REFERENCES

- [1] C. Bizer, R. Cyganiak, S. Auer, and G. Kobilarov. Dbpedia. org—querying wikipedia like a database. In *Proc. Int. World Wide Web Conf*, 2007.
- [2] S. Calegari and G. Pasi. Personal ontologies: Generation of user profiles based on the yago ontology. *Information Processing & Management*, 2012.
- [3] M. Daoud, L.-T. Lechani, and M. Boughanem. Towards a graph-based user profile modeling for a session-based personalized search. *Knowledge and Information Systems*, 21(3):365–398, 2009.
- [4] M. Daoud, L. Tamine, and M. Boughanem. A personalized graph-based document ranking model using a semantic user profile. In *User Modeling, Adaptation, and Personalization*, pages 171–182. Springer, 2010.
- [5] W. M. Darling, M. J. Paul, and F. Song. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. *EACL 2012*, 2012.
- [6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- [7] J. Gantz and D. Reinsel. Extracting value from chaos. *IDC iView*, pages 1–12, 2011.
- [8] Y. Labrou and T. Finin. Yahoo! as an ontology: using yahoo! categories to describe documents. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 180–187. ACM, 1999.
- [9] Y. Li and N. Zhong. Mining ontology for automatically acquiring web user information needs. *Knowledge and Data Engineering, IEEE Transactions on*, 18(4):554–568, 2006.
- [10] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, pages 1–137, 2011.
- [11] G. A. Miller et al. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [12] R. Mizoguchi. Part 3: Advanced course of ontological engineering. *New Generation Computing*, 22(2):193–220, 2004.
- [13] P. Moravec, M. Kolovrat, and V. Snasel. Lsi vs. wordnet ontology in dimension reduction for information retrieval. *Databases, Texts*, page 18, 2004.
- [14] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.
- [15] R. Navigli. A quick tour of word sense disambiguation, induction and related approaches. *SOFSEM 2012: Theory and Practice of Computer Science*, pages 115–129, 2012.
- [16] E. Sanchez. *Fuzzy Logic and the Semantic Web*, volume 1. Elsevier Science, 2006.
- [17] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534. ACM, 2007.
- [18] M. Speretta and S. Gauch. Miology: A web application for organizing personal domain ontologies. In *Information, Process, and Knowledge Management, 2009. eKNOW'09. International Conference on*, pages 159–161. IEEE, 2009.
- [19] Z. Su, J. Yan, H. Ling, and H. Chen. Research on personalized recommendation algorithm based on ontological user interest model. *Journal of Computational Information Systems*, 8(1):169–181, 2012.
- [20] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.
- [21] X. Tao, Y. Li, and N. Zhong. A personalized ontology model for web information gathering. *Knowledge and Data Engineering, IEEE Transactions on*, 23(4):496–511, 2011.
- [22] L. A. Zadeh. Fuzzy sets, information and control, 8 (3): 338-353, 1965.
- [23] N. Zhong. Representation and construction of ontologies for web intelligence. *International Journal of Foundations of Computer Science*, 13(04):555–570, 2002.