OntoGen: Semi-automatic Ontology Editor

Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic

Department of Knowledge Technologies, Institute Jozef Stefan,
Jamova 39, 1000 Ljubljana, Slovenia
{blaz.fortuna, marko.grobelnik, dunja.mladenic}@ijs.si

Abstract. In this paper we present a semi-automatic ontology editor as implemented in a new version of OntoGen system. The system integrates machine learning and text mining algorithms into an efficient user interface lowering the entry barrier for users who are not professional ontology engineers. The main features of the systems include unsupervised and supervised methods for concept suggestion and concept naming, as well as ontology and concept visualization. The system was tested in extensive user trails and in several real-world scenarios with very positive results.

Keywords: Ontology Editor, Ontology Learning, Text Mining, Machine Learning.

1 Introduction

The rapid growth of documents, web pages and other types of textual content pose a great challenge to modern content management systems. Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier access to the information for the users. However, all major ontology editors (such as Protégé[3], OntoStudio[4], …) are fully manual and offer little support to the users for structuring domains.

OntoGen [1, 2] is a semi-automatic and data-driven ontology editor focusing on editing of topic ontologies (a set of topics connected with different types of relations). The system combines text-mining techniques with an efficient user interface to reduce both: the time spent and complexity for the user. In this way it bridges the gap between complex ontology editing tools and the domain experts who are constructing the ontology and do not necessarily have the skills of ontology engineering. The two main characteristics of the system are the following.

- **Semi-Automatic** – The system is an interactive tool that aids the user during the ontology construction process. It suggests: concepts, relations between the concepts, names for the concepts, automatically assigns instances to the concepts, visualizes instances within a concept and provides a good overview of the ontology to the user through concept browsing and various kind of visualization. At the same time the user is always in full control of the systems actions and can fully adjust all the properties of the ontology by accepting or rejecting the system's suggestions or manually adjusting them. This lets the user to establish a trust

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part II, HCII 2007, LNCS 4558, pp. 309–318, 2007.

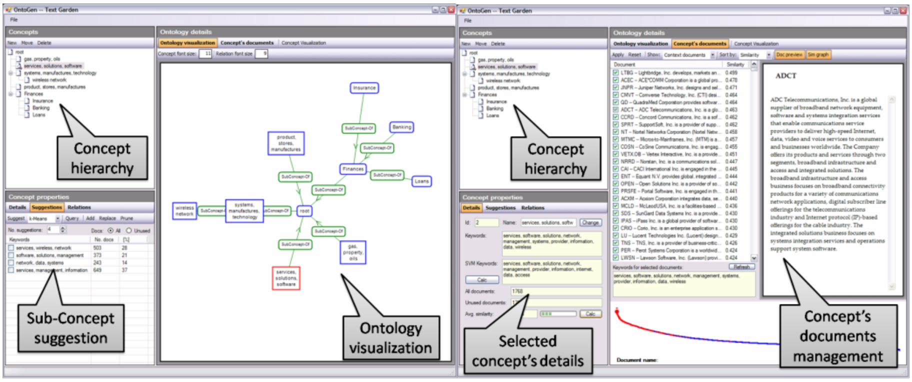
towards the system in a way that he has a full control over all the modifications to the edited ontology.

- **Data-Driven** – Most of the aid provided by the system is based on the underlying data provided by the user typically at the beginning of the ontology construction. The data reflects the structure of the domain for which the user is building ontology. The data is provided as a document corpus where ontological instances are either the documents themselves or name-entities occurring in the documents. The system supports automatic extraction of instances (used for learning concepts) and co-occurrences of instances (used for learning relations between the concepts) from the data.

The rest of the paper is structured as follows. In Section 2 we provide a general overview of OntoGen system, following with detailed description of implementation of the major features in Section 3. We conclude the paper with description of user trails (Section 4).

2 Overview of OntoGen System

Major features of the system generally serve one or both of the two major design goals of OntoGen: (1) visualization and exploration of existing concepts from the ontology and (2) addition of new concepts or modification of existing concept trough simple straightforward procedures aided by machine learning, text mining.



No. suggestions: 4 Docs: ● All ○ Unused

Keywords	No. docs	[%]
□ services, wireless, network	503	28
□ software, solutions, management	373	21
□ network, data, systems	243	14
□ services, management, information	649	37

Id: 2 Name: services, solutions, softw

Keywords: services, software, solutions, network, management, systems, provider, information, data, wireless

SVM Keywords: services, software, solutions, network, management, provider, information, internet, data, access

All documents: 1768

Document	Similarity
☑ LTBG – Lightbridge, Inc. develops, markets an...	0.499
☑ ACEC – ACE*COMM Corporation is a global pro...	0.478
☑ JNPR – Juniper Networks, Inc. designs and sel...	0.471
☑ CMVT – Converse Technology, Inc. (CTI) desi...	0.464
☑ QD – QuadraMed Corporation provides softwar...	0.464
☑ ADCT – ADC Telecommunications, Inc. is a glo...	0.463
☑ CCRD – Concord Communications, Inc. is a sof...	0.462
☑ SPRT – SupportSoft, Inc. is a provider of supp...	0.462
☑ NT – Nortel Networks Corporation (Nortel Netw...	0.458
☑ MTMC – Micros-to-Mainframes, Inc. (MTM) is a...	0.457
☑ COSN – CoSine Communications, Inc. is engag...	0.455
☑ VETX.OB – Vertex Interactive, Inc. is a provide...	0.451
☑ NRRD – Norstan, Inc. is a communications sol...	0.447
☑ CAI – CACI International Inc. is engaged in the ...	0.445
☑ ENT – Equant N.V. provides global, integrated ...	0.444
☑ OPEN – Open Solutions Inc. is a provider of so...	0.442
☑ PRSFE – Portal Software, Inc. is engaged in th...	0.441
☑ ACXM – Acxiom Corporation integrates data, se...	0.440
☑ MCLD – McLeodUSA, Inc. is a facilities-based ...	0.436
☑ SDS – SunGard Data Systems Inc. is a provide...	0.430
☑ IPAS – iPass Inc. is a global provider of softwar...	0.430
☑ CRIO – Corio, Inc. is an enterprise application s...	0.430
☑ LU – Lucent Technologies Inc. (Lucent) design...	0.429
☑ TNS – TNS, Inc. is a provider of business-critic...	0.426
☑ PER – Perot Systems Corporation is a worldwid...	0.424
☑ LWSN – Lawson Software, Inc. (Lawson) provi...	0.424

Keywords for selected documents: services, software, solutions, network, management, systems, provider, information, data, wireless

ADCT

ADC Telecommunications, Inc. is a global supplier of broadband network equipment, software and systems integration services that enable communications service providers to deliver high-speed Internet, data, video and voice services to consumers and businesses worldwide. The Company offers its products and services through two segments, broadband infrastructure and access and integrated solutions. The broadband infrastructure and access business focuses on broadband connectivity products for a variety of communications network applications, digital subscriber line offerings for the telecommunications industry and Internet protocol (IP)-based offerings for the cable industry. The integrated solutions business focuses on systems integration services and operations support system software.

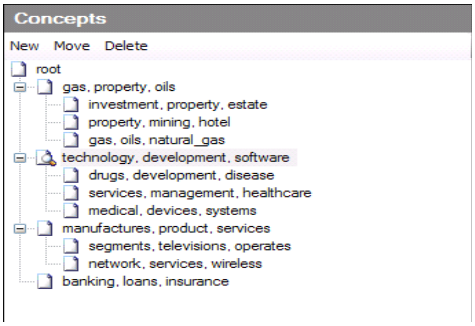
Fig. 1. *Left*: the user is getting suggestions for the sub-concepts of the selected concept (left bottom part); the ontology is visualized as a concept hierarchy in textual mode (left upper part) and in graphical mode (right central part). *Right*: the user is investigating the selected concept by browsing through the concept statistics (bottom left), the concept's documents including their content (upper right) and the graph of similarities of the documents towards the concept (bottom right).

The main window of the system provides multiple views on the ontology. A tree-view on the ontology, as it is intuitive for most users, presents a natural way to represent a concept hierarchy. This view is exposed as a standard Windows control used to show folder structure and as a visualization offering a one-glance view of the whole ontology. Each concept from the ontology is further exposed by the set of the most informative keywords for the target concept being automatically extracted using text-mining techniques [1] and through a topic-map of the documents belonging to the concept (more details in Section 3). Figure 1 shows the standard layout of the system.

Concept suggestions play a central part in the system. We provide unsupervised and supervised methods for generating suggestions. **Unsupervised** learning methods automatically generate a list of sub-concepts for a currently selected concept by using k-means clustering and latent semantic indexing (LSI) techniques [1, 8] to generate a list of possible sub-concepts. **Supervised** learning methods on the other hand require the user to have a rough idea about a new topic – this is identified through a query returning the documents. The system automatically identifies the documents that correspond to the topic and the selection can be further refined by the user-computer interaction through an active learning loop [1, 10] using a machine learning technique for semi-automatic acquisition of the user knowledge. The system also provides means of defining concept by selecting areas in the corresponding topic-map.

3 Implementation of Main Features

In this section we describe the main features of OntoGen including concept hierarchy visualization, management of a concept providing the concept details and suggestions for the concept naming, formation of a new concept based on unsupervised and supervised machine learning methods, concept visualization. We also provide a brief description of some additional features that enable collaborative editing of ontologies and user profiling.



- root
  - gas, property, oils
    - investment, property, estate
    - property, mining, hotel
    - gas, oils, natural_gas
  - technology, development, software
    - drugs, development, disease
    - services, management, healthcare
    - medical, devices, systems
  - manufactures, product, services
    - segments, televisions, operates
    - network, services, wireless
  - banking, loans, insurance

Fig. 2. Tree-view visualization. The user can reposition an existing concept to a different position in the hierarchy or delete an existing concept from the ontology. When moving a concept, the program pops up a dialog window asking the user to select the destination concept.

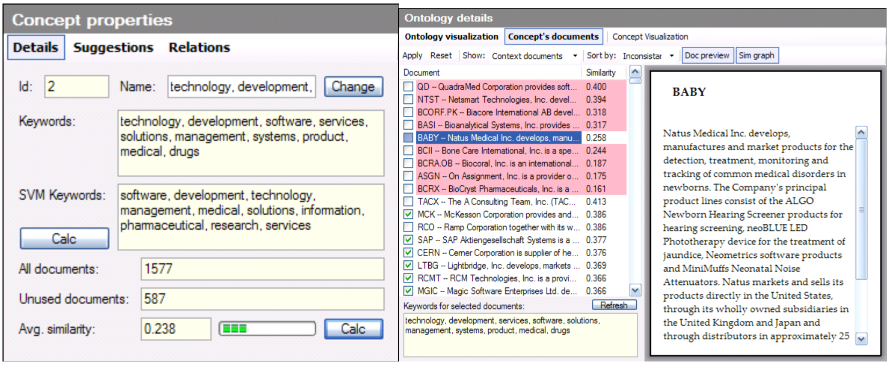
3.1 Concept Hierarchy Visualization

The part of the main window which is always visible to the user is the tree-view visualization of the concept hierarchy (the upper left in Figure 1 “Concept hierarchy”, isolated in Figure 2). It offers a quick overview of all the concepts with their position in the concept hierarchy, direct access to commands for manipulating the hierarchy.

Ontology is also visualized in real time on the right side of the main window ("Ontology visualization" in Figure 1). The root concept is displayed in the centre of the visualization; sub-concepts of the root concept are displayed in the circle around the root concept, and so on. The user can select a concept by clicking on it on the visualization. The currently selected concept is drawn in red color, other concepts are drawn in blue color and the relations are drawn with green color.

3.2 Concept Details and Management

Detail and properties of the selected concept are presented in the bottom left part of the main window ("Selected concept's details" in Figure 1 and left side of Figure 3). In this part the user can change the name of the selected concept, check the concept's main keywords and see the number of instances in the concept (*All documents*), number of instances not in any of the sub-concepts (*Unused documents*) and average similarity of the documents within the concept (*Avg. Similarity*).



Id: 2

Name: technology, development,

Keywords: technology, development, software, services, solutions, management, systems, product, medical, drugs

SVM Keywords: software, development, technology, management, medical, solutions, information, pharmaceutical, research, services

All documents: 1577

Unused documents: 587

Avg. similarity: 0.238

Show: Context documents

Sort by: Inconsistar

Document	Similarity
□ QD -- QuadraMed Corporation provides soft...	0.400
□ NTST -- Netsmart Technologies, Inc. devel...	0.394
□ BCORF.PK -- Biacore International AB devel...	0.318
□ BASI -- Bioanalytical Systems, Inc. provides ...	0.317
□ BABY -- Natus Medical Inc. develops, manu...	0.258
□ BCII -- Bone Care International, Inc. is a spe...	0.244
□ BCRA.OB -- Biocoral, Inc. is an international...	0.187
□ ASGN -- On Assignment, Inc. is a provider o...	0.175
□ BCRX -- BioCryst Pharmaceuticals, Inc. is a ...	0.161
□ TACX -- The A Consulting Team, Inc. (TAC...	0.413
☑ MCK -- McKesson Corporation provides and...	0.386
□ RCO -- Ramp Corporation together with its w...	0.386
☑ SAP -- SAP Aktiengesellschaft Systems is a ...	0.377
☑ CERN -- Cerner Corporation is supplier of he...	0.376
☑ LTBG -- Lightbridge, Inc. develops, markets ...	0.369
☑ RCMT -- RCM Technologies, Inc. is a provi...	0.366
☑ MGIC -- Magic Software Enterprises Ltd. de...	0.366

Keywords for selected documents: technology, development, services, software, solutions, management, systems, product, medical, drugs

BABY

Natus Medical Inc. develops, manufactures and market products for the detection, treatment, monitoring and tracking of common medical disorders in newborns. The Company's principal product lines consist of the ALGO Newborn Hearing Screener products for hearing screening, neoBLUE LED Phototherapy device for the treatment of jaundice, Neometrics software products and MiniMuffs Neonatal Noise Attenuators. Natus markets and sells its products directly in the United States, through its wholly owned subsidiaries in the United Kingdom and Japan and through distributors in approximately 25

Fig. 3. *Left*: component showing details of the currently selected concept. *Right*: Interface for managing concepts.

There are two keyword extraction methods implemented in the system [1]. The first one, shown under *Keywords*, is composed from words most descriptive for the content of the concept's instances and the second one, shown under *SVM Keyword*, is composed from words most distinctive for the selected concept with regards to its sibling concepts in the hierarchy. Since SVM keywords are expensive to extract they are computed only on user's request (with button *"Calc"*).

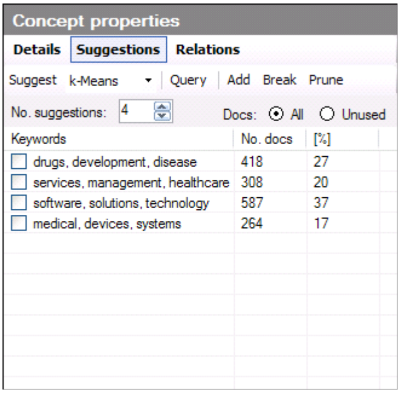
After a new concept is added to the ontology, either with unsupervised or supervised methods, the system automatically assigns instances to it. Since this assignment is often not perfect the user can manually reassign instances to and from the concept through the interface shown in right part of Figure 3. Interface is composed of the list of instances where the instances from the concept are checked. Assignment of instances can be simply changed by checking or un-checking them. The content of the selected instance is displayed on the right side. OntoGen also has functionality for detecting specific inconsistencies inside ontology. When an instance does not belong to the selected concept but is in one of its sub-concepts, then it has highlighted using red background (for example check Figure 3).

3.3 Concept Suggestions

One of the main parts of the system is suggestions of possible new concept to add to the ontology. There are two different approaches implemented for concept learning. In the *unsupervised* approach the system provides suggestions for possible sub-concepts of the selected concept. In the *supervised* approach the user has an initial idea of what a sub-concept should be about and enters it into the system as a query.

3.3.1 Unsupervised

Unsupervised suggestions are based on clustering methods from text-mining [1] and clusters of instances from selected concept are treated as sub-concept suggestions. The user can supervise the number of clusters the clustering methods should return and on which instances the clustering should be performed (*All* for all instances in the concept and *Unused* for instances that are currently not in any of the concept's sub-concepts). An example of suggestions is given in Figure 4. User can accept and add to the ontology any of the provided suggestions by simply checking it and clicking *Add* button.



Suggest k-Means

No. suggestions: 4

Docs: ● All ○ Unused

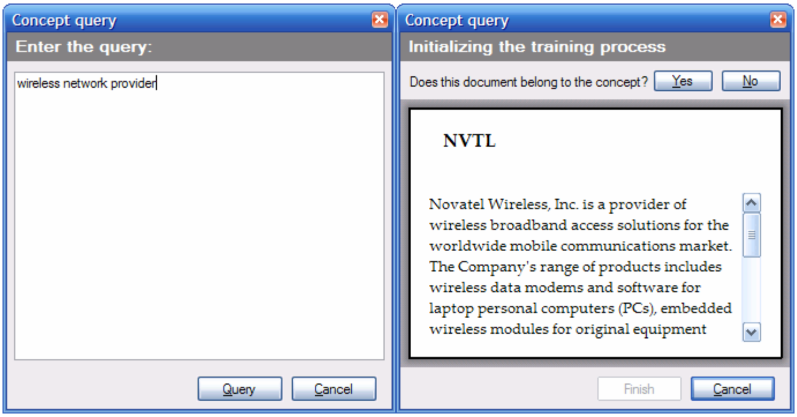
Keywords	No. docs	[%]
□ drugs, development, disease	418	27
□ services, management, healthcare	308	20
□ software, solutions, technology	587	37
□ medical, devices, systems	264	17

Fig. 4. The clustering algorithm prepares suggestions and the program displays them in the list (see lower half). Each suggested sub-concept is described by the extracted main keywords (using the centroid method) and the size. A longer list of keywords is displayed when the user moves with mouse over the suggestion.

3.3.2 Supervised

A new feature in OntoGen is a supervised method for adding concepts. It is based on SVM active learning method [2] and is only applied to the instances from the selected concept. The user supervision is provided first by a query or a set of keywords describing the concept that the user has in mind and followed by a sequence of questions (see figure 5).

On each step the system asks if a particular instance belongs to the concept and the user can select *Yes* or *No*. The questions are chosen from the instances on the border between being relevant to the query or not and are therefore most informative to the system. The system refines the suggested concept after each reply from the user and the user can decide when to stop the process based on how satisfied he is with the suggestions. After the concept is constructed it is added to the ontology as a sub-concept of the selected concept.



wireless network provider

Does this document belong to the concept?

NVTL

Novatel Wireless, Inc. is a provider of wireless broadband access solutions for the worldwide mobile communications market. The Company's range of products includes wireless data modems and software for laptop personal computers (PCs), embedded wireless modules for original equipment

Fig. 5. *Left*: user entering a query. *Right*: question example from the system for the given query. Instance and its description is shown to aid the user at the decision.

3.3.3 Example of Unsupervised and Supervised Concept Suggestion

There is a fundamental difference between the unsupervised and supervised methods. The main advantage of unsupervised method is that it requires very little input from the user (number of clusters). The unsupervised method provide well balanced suggestions for sub-concepts based on the instances and can be very useful for exploring the data.

The supervised method on the other hand requires more input. The user has to first figure out what should the sub-concept be, he has to describe the sub-concept trough a query and go through the sequence of questions to clarify the query. This is intended for the cases where the user has a clear idea of the sub-concept he wants to add to the ontology but the unsupervised methods do not discover it. Idea for a query can come from concept visualization presented in the next section.

3.4 Concept Visualization

Instances of selected concept can be visualized using text-document visualization techniques described in [6]. The instances are presented as points on a map in such a way that each instance located close to similar instance and far from less-similar instances. The similarity between instances is calculated based on their textual description using *cosine similarity* (well know similarity measure from text-mining and information retrieval). The density of instances in specific part of the map is presented by the background texture. Most common keywords are shown for each area of the map and when the user moves the mouse around the map an extended list of the most common keywords is shown for the area around the mouse. The user can also zoom-in to see specific areas in details.

The visualization is tightly integrated into the system. The user can add new sub-concepts to the ontology by simply selecting a group of instances from the map. This is particularly useful since visualization often represents the groupings of similar instances into clusters. There is an example of visualization in Figure 6.

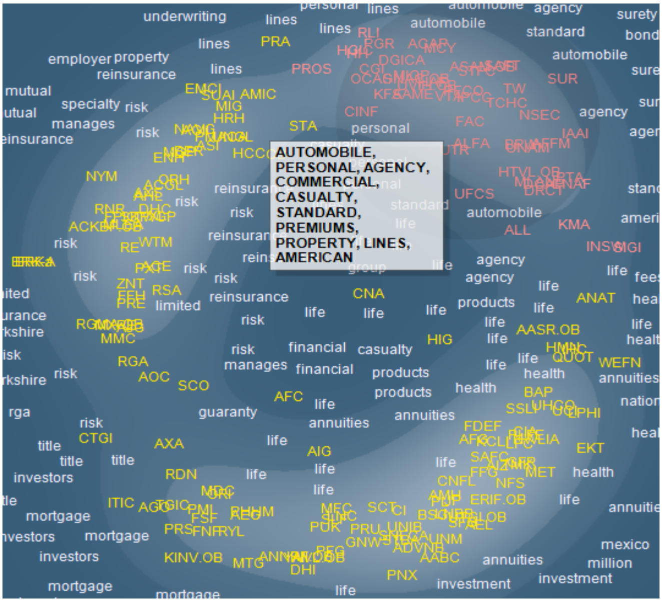


Fig. 6. The visualization is showing a map where each instance is a description of one insurance company. We can see that instances nicely group into three distinct clusters: reassurance, home and car insurance and life insurance. The user selected the instances from home and car insurance using mouse (the pink instances) to add them as a sub-concept to the ontology.

3.5 Collaborative Editing and User Profiling

The system also offers collaborative editing of ontologies where the user can use topics and relations which were previously constructed by other users. The system supports the user by suggesting similar topics/relations from the collection of ontologies.

User profiling is also used to tune the human ontology interaction based on the previous work of the users. This is done by recording previous choices that the user made when constructing ontologies and using them as an extra input to provide a personalized view on the underlying data and the ontology constructed so far trough "personalized word weighting schemas" (instead of using predefined schemas such as TFIDF [7]).

4 Evaluation

The first prototype of the system [1] was already successfully applied in case studies of several commercial projects from the domain of business, legislation and digital libraries. The users participating in the case studies were domain experts with limited experience in ontology construction. The feedback we got from the user was positive and we have used it to further improve the user interface. For example, simple method for moving concepts in the hierarchy (Section 3.1) is based on the feedback from users. In particular, the system enabled the users to model ontologies which would be significantly more difficult and expensive to model otherwise.

4.1 User Trail Setup

The latest version of the system [2] was extensively tested trough user trails at Faculty of Arts and Sciences in Rijeka [11]. The general goal of user trials was to gain as much useful information as it could be done in order to objectively assess the current stage of development of the OntoGen and its functionalities [9].

For that purpose two different groups of users were chosen to provide the useful data from two different points of view; one group consisted of 48 Computer sciences students and other from 43 Psychology and Pedagogy sciences students. That kind of user selection enabled the gathering of data from subjects who, generally, have good computer skills, only one group has more than average informational knowledge and understands the technical background of software development, and other group has more than common knowledge about cognitive processes of mental data structuring and organizing.

An average length of user trail session took 90 minutes and consisted of three phases. The first phase was used to introduce the users to the purpose of the task and to explain the procedures which will follow in the second and third phases of the trail. Main data collection is second phase of user trials and it had three main parts – before, during and after the demonstration of OntoGen system. In these parts we recorded several measure points for data collection:

1. Filling in the Initial questionnaire
2. In the second part the students tested the ontology construction system OntoGen. After demonstration of the OntoGen system by the facilitator, they were asked to fulfill the following two tasks:
   - Construct ontology which captures the areas covered by the companies in the given collection of descriptions of 7177 companies.
   - Construct two different ontologies on top of the same collection of 5000 news articles from Reuters news agency. The first ontology should group instances based on geographical properties, and the second based on topic of news articles.
3. In the third part, through the End questionnaire, they were asked about their experiences with the tested system. After the trial students also had a possibility to discuss their experiences with other participants involved in trials.

Conclusion is the third phase of user trial in which it is necessary to thank all the users for participating in user trial. Also, at this moment, it is very important to discus, if necessary, any question or thought aroused from the tasks during the trials.

4.2 User Trail Results

The general impression of OntoGen tool is marked “mostly good” to “very good” on scale from “not good at all”, “sufficient”, “mostly good”, “very good” to “excellent”. The great majority of users think that the tool is useful, but there is some resentment in content presentation of the main areas; the overall mark would be “mostly good” for Ontology visualisation, Concept tree and Concept Properties. The similar situation occurred in assessment of graphical interface; overall impression is “mostly good”, but there were objections expressed concerning the attractiveness. Users repeatedly pointed out that there should be more colours.

Qualitative analysis revealed that one of the main advantages of OntoGen is managing large data bases with easiness. According to users, the tool is efficient, saves time and effort and gives a lot of space for user intervention. Disadvantages, as perceived, mostly concern unattractive look, abstract conception and occasional slowness and need to learn how to use it first.

Users also think that there is a clear distinction between two ways of ontology creation (unsupervised and supervised). Here the users only draw attention to the improvement of the graphical layout. Lack of detailed instructions and need for all the options in one place were pointed out as the main difficulties in generating concepts.

Visualization of concepts proved to be a great help in choosing sub-concepts for majority of users. Most of the users found the current state of visualization appropriate, but in need of more attractive graphical solutions and even bigger window or perhaps larger monitors. Furthermore, users think that with adding of concepts, picture becomes cluttered which makes it hard to inspect.

When it comes to concept documents management, users think that it is quite good manageable. There are no significant differences in assessing that part of OntoGen; general mark for all the operations included in document management within the concepts is “very good”.

5 Conclusions

In this paper we presented an extension of ontology editors based on extensive use of machine learning and text mining methods which aid the user in the major steps of ontology construction process. The inherent danger of such approach is than an introduction of new text mining methods could make the whole system hard to comprehend and use by the domain experts, which in general do not have any knowledge of text mining. However, our user trails showed that tool OntoGen managed to avoid such missteps and that it can be used by people without background from machine learning.

As part of future work we plan to use the comments of the user trails to improve the user interface and user interactions. We also plan to extend the system with support for user collaboration when editing larger ontologies and to support for reuse of ontologies or parts of ontologies produced by other users.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP), NeOn Lifecycle Support for Networked Ontologies (IST-4-027595-IP) and PASCAL Network of Excellence (IST-2002-506778).

References

1. Fortuna, B., Mladenić, D., Grobelnik, M.: Semi-automatic construction of topic ontologies. In: Ackermann, M., Berendt, B., Grobelnik, M., Hotho, A., Mladenič, D., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., van Someren, M. (eds.) Semantics, Web and Mining. LNCS (LNAI), vol. 4289, pp. 121–131. Springer, Heidelberg (2006)
2. Fortuna, B., Grobelnik, M., Mladenic, D.: Semi-automatic Data-driven Ontology Construction System. In: Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia (2006)
3. More information on Protégé (Stanford University) and the download of the latest version is available from http://protege.stanford.edu
4. Information about OntoStudio (Ontoprise GmbH) is available through http://www.ontoprise.de/content/e1171/e1249/index_eng.html
5. Fortuna, B., Grobelnik, M., Mladenic, D.: Visualization of Text Document Corpus. Informatica 29, 497–502 (2005)
6. Fortuna, B., Grobelnik, M., Mladenic, D.: Background Knowledge for Ontology Construction. In: Proceedings of the 15th International World Wide Web Conference WWW 2006, Edinburgh, Scotland (May 23-26, 2006)
7. Manning, C.D., Schutze, H.: Foundations of statistical Natural Language Processing, MIT Press.
8. Mladenic, D., Grobelnik, M.: Evaluation of ontology generation for case studies, Technical Report IJS-DP, J. Stefan Institute, Ljubljana (January 2007)
9. Tong, S., Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. In: Proceedings of 17th International Conference on Machine Learning
10. Ilijasic Misic, I., Taksic, V., Kovacic, B., Mohoric, T.: Design of a user study and evaluation of software tool OntoGen V 2.0, Technical Report FAS Rijeka (December 2006)