USING QUERY EXTENSION AND USER FEEDBACK TO IMPROVE PUBMED SEARCH

Viktoria Dorfer [(a)], Sophie A. Blank [(b)], Stephan M. Winkler [(c)], Thomas Kern [(d)], Gerald Petz [(e)], and Patrizia Faschang [(f)]

[(a,b,c,d)] University of Applied Sciences Upper Austria
School of Informatics, Communications and Media, Bioinformatics Research Group
Softwarepark 11, 4232 Hagenberg, Austria

[(e,f)] University of Applied Sciences Upper Austria
School of Management, Digital Economy Research Group
Wehrgrabengasse 1-3, 4400 Steyr, Austria

[(a)] viktoria.dorfer@fh-hagenberg.at, [(b)] sophieanna.blank@students.fh-hagenberg.at, [(c)] stephan.winkler@fh-hagenberg.at, [(d)] thomas.kern@fh-hagenberg.at, [(e)] gerald.petz@fh-steyr.at, [(f)] patrizia.faschang@fh-steyr.at

ABSTRACT

PubMed is a search engine that is widely used to search for medical publications. A common challenge in information retrieval, and thus also when using PubMed, is that broad search queries often result in lists of thousands of papers that are presented to the user, too narrow ones often yield small or even empty lists. To address this problem we here present a new PubMed search interface with query extension using keyword clusters generated with evolutionary algorithms to obtain more specific search results. Users can choose to add various words to their query and then rate search results; this scoring is stored in a database to enable learning from user feedback to improve keyword cluster optimization as well as query extensions. We show how users can extend PubMed queries using previously generated keyword clusters, rate query results, and use these ratings for optimizing parameters of the keyword clustering algorithms.

Keywords: Bioinformatical Information Retrieval, PubMed, Query Extension, Keyword Clustering

1. INTRODUCTION

Search queries in PubMed (PubMed 2011) often return a huge number of papers, many of them not relevant for the user. To address this problem of irrelevance, information retrieval provides a suitable solution, namely query extension. User queries are extended with matching terms to result in fewer, but relevant search results. This method is based on clusters of keywords and / or documents; from these clusters words are selected that belong to the same clusters as the search terms defined by the user, and thus queries can be extended automatically. Figure 1 shows a schematic view of an exemplary keyword clustering solution consisting of three clusters.

In this paper we present a new interface for PubMed search that uses various keyword clusters to extend user queries. These keyword clusters have been generated using various evolutionary algorithms as described in Dorfer, Winkler, Kern, Petz, and Faschang (2010) and Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011). User ratings of the query results (returned by PubMed for queries that were extended using keyword clusters) will then be used to improve the generation of keyword clusters and query extensions.

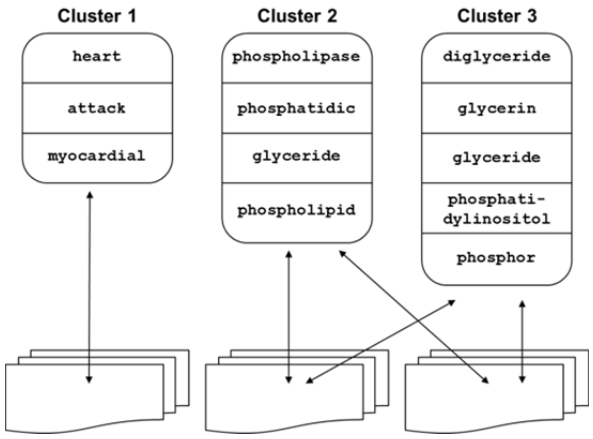


Figure 1: Example of a keyword clustering solution

2. A NEW PUBMED SEARCH INTERFACE

We have developed a PubMed search interface that uses previously generated keyword clusters to suggest words that fit the terms the user entered as query. Several clustering files can be selected to search for suitable extensions.

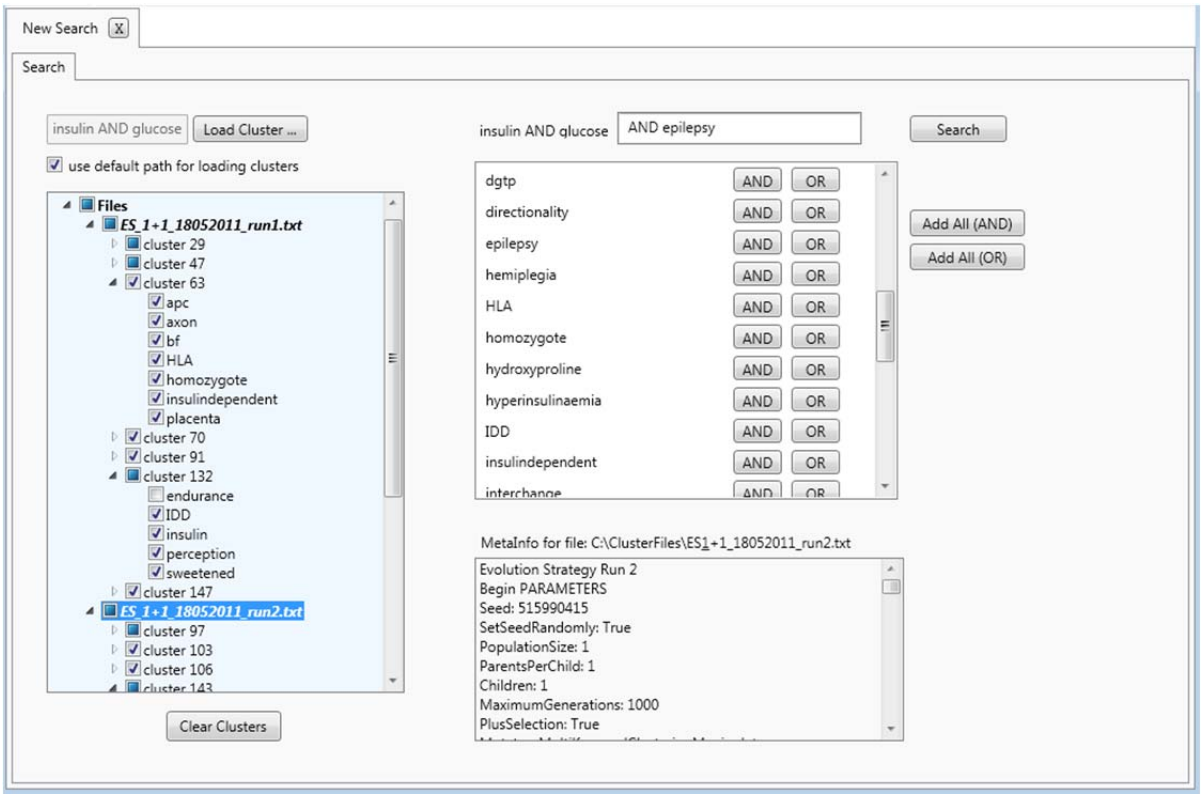


Figure 2: PubMed query extension using previously generated keyword clusters

In Figure 2 a screenshot of the search and extension interface is provided. On the left hand side the keyword clusters containing terms or parts of terms of the user query are shown; these keywords can be selected for query extension.

On the right side the chosen terms can be added to the query using ADD or OR conjunctions. More than one search query can be submitted in parallel to be able to compare results with different selections of keywords extending the original query.

3. KEYWORD CLUSTER GENERATION

As mentioned before, these keyword clusters have been generated using evolutionary algorithms, driven by mutation, crossover and selection (see Dorfer, Winkler, Kern, Petz, and Faschang (2010) for further details on the operators implemented for this problem class) using a fitness function that is used for evaluating solution candidates. This fitness function takes various features into account we consider important for keyword clustering solution candidates and is defined by the following equation (Dorfer, Winkler, Kern, Petz, and Faschang (2010)):

$$F = \alpha \cdot A + \beta \cdot B + \gamma \cdot C + \delta \cdot D + \varepsilon \cdot E + \zeta \cdot G\,. \quad (1)$$

Parameter A represents the number of clusters a document is assigned to; parameter B the data coverage; C – the cluster confidence – quantifies to which amount the keywords of the cluster are also present in the documents assigned to the cluster, whereas in parameter D – the document confidence – the amount of keywords in the documents also present in the cluster is represented. Parameter E regards the distribution of the documents in the clusters and G considers the number of generated clusters with respect to the data set size. α, β, γ, δ, ε, and ζ are weighting factors that are necessary to be able to emphasize or neglect specific features, depending on the user's needs.

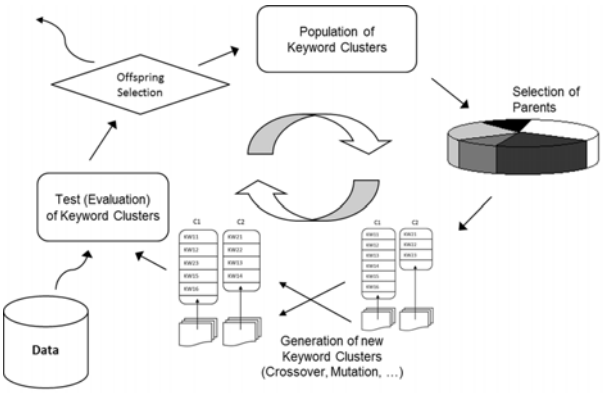


Figure 3: Genetic algorithm with offspring selection optimizing keyword clusters

In Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011) we have presented a detailed comparison of the performance of various evolutionary algorithms solving this keyword clustering problem (KCP), including the evolution strategy (ES) (Schwefel 1994), the genetic algorithm (GA) (Holland 1975), the

genetic algorithm with offspring selection (OSGA) (Affenzeller, Winkler, Wagner, and Beham 2009), and the elitist non-dominated sorting genetic algorithm (NSGA-II) (Deb, Pratap, Agarwal, and Meyarivan 2002). In Figure 3 we show the most important parts of an evolutionary algorithm solving the KCP (in particular, an OSGA is depicted).

The evolutionary process in ESs is mainly based on mutation and selection, whereas evolution in GAs is directed by the interaction of mutation, crossover and selection; the OSGA is an extension of the standard GA and includes an additional offspring selection step. The NSGA-II is a multi-objective approach, which optimizes various objectives in parallel, as we have identified parameters B, C, and D as potentially contrary.

We have shown that intense parameter settings and weighting factor tuning is necessary to receive good solutions using single-objective algorithms and that ESs and the NSGA-II perform best in terms of achievable solution quality (Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011)).

The data set used to generate the mentioned keyword clusters has been taken from the 9th Text Retrieval Conference (TREC-9) from the year 2000 (Vorhees and Harman 2000), consisting of 36,890 samples based on PubMed entries including title, abstract, authors, its medical subject headings (MeSH), publication type and source.

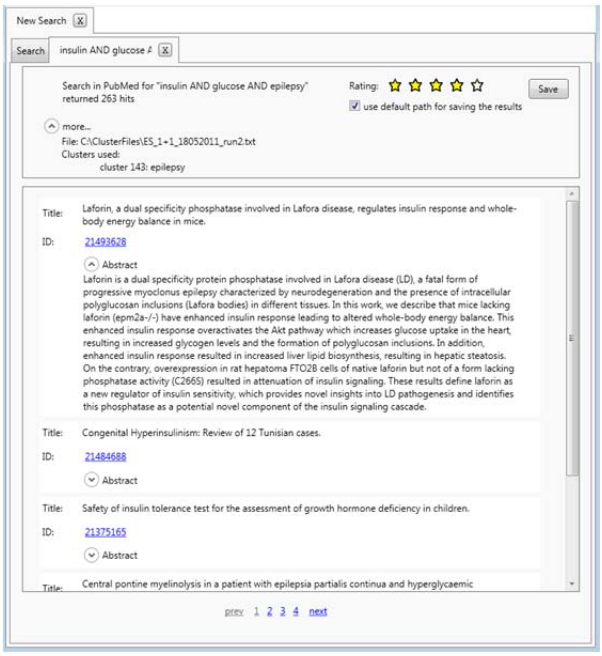


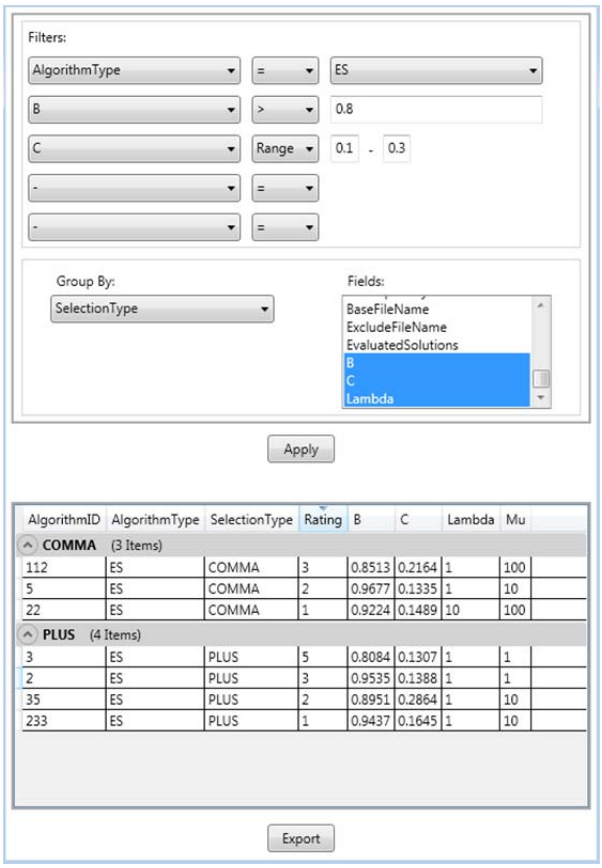
Figure 4: Example of search results with rating possibility

4. USER FEEDBACK

After the search has been performed the user can rate the results (see Figure 4); this rating in combination with the words used for query extension and the algorithm parameters the keyword clusters have been created with (algorithm type, algorithm parameters, documents used for clustering, …) are stored in a database in order to be able to examine the settings and algorithms which delivered the best results.

Another interesting aspect in the context of biomedical information retrieval is whether there are specific keywords that, when included in the query, never lead to satisfying solutions. By analyzing user feedback these words can be identified and collected in a blacklist to warn the user as soon as such a term is included in the query.

User feedback cannot only be used for the generation of a blacklist, but also to improve the parameter settings of the keyword clustering algorithms. This shall enable an iterative improvement of parameter settings leading to better and more suitable clusters which can then be used in query extension.



AlgorithmID	AlgorithmType	SelectionType	Rating	B	C	Lambda	Mu
COMMA (3 Items)							
112	ES	COMMA	3	0.8513	0.2164	1	100
5	ES	COMMA	2	0.9677	0.1335	1	10
22	ES	COMMA	1	0.9224	0.1489	10	100
PLUS (4 Items)							
3	ES	PLUS	5	0.8084	0.1307	1	1
2	ES	PLUS	3	0.9535	0.1388	1	1
35	ES	PLUS	2	0.8951	0.2864	1	10
233	ES	PLUS	1	0.9437	0.1645	1	10

Figure 5: Analysis of algorithm parameters that were applied for forming keyword clusters using various evolutionary algorithms. The keyword clusters have been used for query extension and the search results have been rated; algorithm settings are ranked according to the user's feedback

In Figure 5 we show an example of the examination of rating results and algorithm specific parameter settings to find out whether these settings yield appropriate solutions. All fields of the database can be selected for filtering, including general algorithm parameters (such as algorithm type, population size, and also fitness function specific parameters as – for example – B and D) and of course the user ratings. The previously explained keyword clusters and the evolutionary algorithms used to generate these clusters provide the basis for this analysis. Based on the results of these examinations we want to further improve the algorithm parameters to optimize the generation of keyword clusters; this shall extend our optimization algorithm with a user-driven feedback component.

5. APPLICATION EXAMPLES

Using the best keyword clustering files identified in Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011), generated by ES, GA, OSGA and the NSGA-II, a detailed analysis on the differences of the generated clusters and on the performance of the different algorithms with the different parameter settings with respect to user feedback can now be performed.

In Figure 6 we provide an application example, in particular, we search for "urethra". We can clearly see the differences in the clusters generated by ES and NSGA-II. As evolution in evolution strategies solving the KCP is dependent on the fitness function given in Equation 1, quite different keywords are grouped together in contrast to the clusters generated by the NSGA-II, which in this example optimizes only parameters B, C, and D. On the right bottom of the window detailed information on the used algorithm and its parameter settings to generate the cluster is given. All shown keywords can be chosen for query extension, optionally in combination; this selection does not depend on the algorithm the clusters have been generated by.

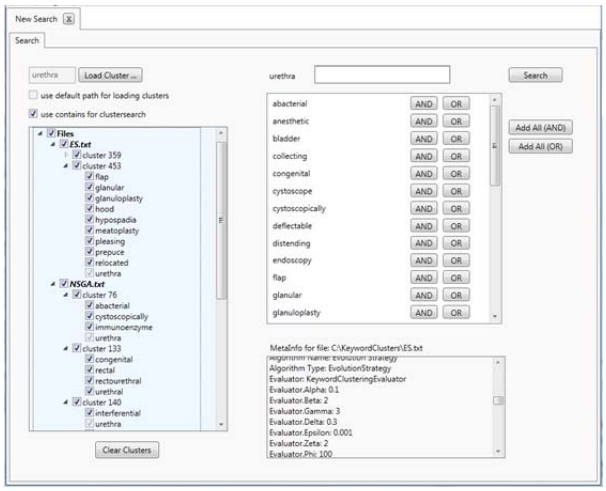


Figure 6: Application example using keyword clusters generated by two different evolutionary algorithms

After users have performed various queries and have rated the obtained results accordingly, we can now have a further look on the performance of the various evolutionary algorithms and their corresponding settings in terms of parameters and also weighting factors. In Figure 6 we can see that here keyword clusters generated by evolution strategy with specific weighting factor settings and also generated by the NSGA-II provided clusters the users ranked best. However, this does not have to be the case in all examples; one can assume that specific user needs and specific combinations of terms are more likely to be produced by a genetic algorithm, for example. This and other questions can be analyzed by applying various filters on the feedback data, beside the possibility to gain a quick insight in which algorithm performs best regardless of the specific user needs.

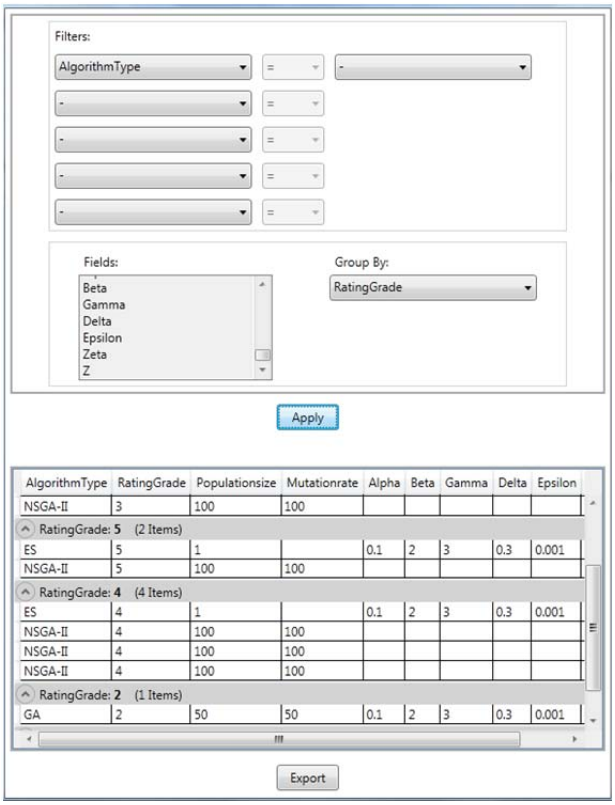


Figure 7: Application example on the analysis of user feedback

As this new search tool shall be used to improve PubMed search results we give another application example here: Assuming – for the sake of simplicity – one wants get some information about the effect of smoking on the lung, a search of these two terms ("smoke AND lung") would yield 7310 hits. In Figure 8 the words that often occur in combination with at least one of these two terms are depicted.

To be sure not to have missed any paper on this topic, the query can be extended with an OR conjunction on the term "pulmonary" to retrieve also all the papers where pulmonary has been consequently used as a synonym for lung, leading to a publication list of 8384 entries.

Alternatively, a reduction of the result space can be obtained by adding more specific terms inspired by the suggestion of the keyword clusters. Adding for example the keyword "asthma", PubMed returns only 699 hits, adding the keyword "haemoptysis" leads to only 6 results.

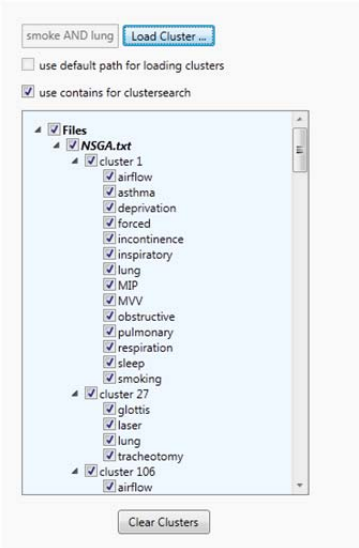


Figure 8: Keyword clusters containing the terms "lung" or "smoke"

6. CONCLUSION

In this paper we have presented a new query interface that used keyword clusters to improve PubMed queries. The used clusters have been generated using various evolutionary algorithms and different parameter settings. In addition to the extension functionality, obtained results can be rated and various analyses can be performed on this kind of user feedback.

We have shown that the various algorithms generate different keyword clusters and that it is therefore reasonable to use several keyword clustering files to choose the words for extension. The provided example described the applicability for extending and narrowing the search space by the use of either synonyms or specifications, depending on the users' needs in the context of PubMed search.

7. FUTURE WORK

We are currently working on an online version for the proposed tool for improved PubMed querying. This version will enable us to address a broader user clientele and then to do further analyses on their feedback. Having now identified the best performing algorithms we plan to slightly adapt parameters and weighting factors and to perform further keyword clustering to hopefully produce even better keyword clusters. This will be an ongoing process as new keyword clusters will lead to new ratings and so on.

As mentioned before we also plan to design a black list of terms or of combinations of terms and algorithms that are less likely to lead to useful search results. We are also working on the automated generation of so-called white lists; these are lists of terms that should always be used in combination to obtain meaningful results. Users can then be warned or encouraged to include specific terms or keyword clusters of specific algorithms in their search.

ACKNOWLEDGMENTS

The work presented in this paper was done within the TSCHECHOW project, sponsored by the basic funding program of the University of Applied Sciences Upper Austria.

REFERENCES

Affenzeller, M., Winkler, S., Wagner, S., Beham, A., 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications.* Chapman & Hall/CRC. ISBN 978-1584886297. 2009.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6:182-197.

Dorfer, V, Winkler, S.M., Kern, T., Blank, S.A., Petz, G., and Faschang, P., 2011. On the Performance of Evolutionary Algorithms in Biomedical Keyword Clustering, *Proceedings of the Genetic and Evolutionary Computation Conference,* 10, July 12-16, 2011, Dublin, Ireland.

Dorfer, V, Winkler, S.M., Kern, T., Petz, G., and Faschang, P., 2010. Optimization of Keyword Grouping in Biomedical Information Retrieval Using Evolutionary Algorithms, *Proceedings of the 22nd European Modeling and Simulation Symposium,* 25-30, October 13-15, 2010, Fes, Morocco.

Holland, J.H., 1975. *Adaption in Natural and Artifical Systems.* University of Michigan Press

PubMed, 2011, Available from: http://www.ncbi.nlm.nih.gov/pubmed

Schwefel, H.-P., 1994. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.* Basel: Birkhäuser Verlag

Vorhees, E.M., Harman, D.K., 2000. *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, Maryland: Department of Commerce, National Institute of Standard and Technology.

AUTHORS BIOGRAPHIES

VIKTORIA DORFER is a senior researcher in the field of Bioinformatics at the Research Center Hagenberg, School of Informatics, Communications and Media. After finishing the diploma degree of bioinformatics in 2007 she was a team member of various projects in the field of bioinformatics and software engineering. She is currently working on biomedical information retrieval within the TSCHECHOW project.

SOPHIE A. BLANK received her bachelor's degree in bioinformatics from the Upper Austria University of Applied Sciences in 2011; she is currently a junior researcher at the Research Center Hagenberg, School of Informatics, Communications and Media working on biomedical information retrieval within the TSCHECHOW project.

STEPHAN M. WINKLER received his MSc in computer science in 2004 and his PhD in engineering sciences in 2008, both from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, bioinformatics, nonlinear model identification, and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the Upper Austria University of Applied Sciences, Campus Hagenberg, and since 2011 he is head of the Bioinformatics Research Group.

THOMAS KERN is head of the Research Center Hagenberg, School of Informatics, Communications and Media, Upper Austria University of Applied Sciences (UAS). He finished his studies in Software Engineering in 1998. After some work experience in industry he started his academic career at UAS in autumn 2000 as research associate and lecturer for algorithms and data structures, technologies for knowledge based systems, semantic systems and information retrieval. Since 2003 he has conducted several application oriented R&D projects in the fields of bioinformatics and software engineering.

GERALD PETZ is director of the degree course Marketing and Electronic Business at the University of Applied Sciences in Upper Austria, School of Management in Steyr. His main research areas are Web 2.0, electronic business and electronic marketing; he has also conducted several R&D projects in these research fields. Before starting his academic career he was project manager and CEO of an internet company.

PATRIZIA FASCHANG received her bachelor's degree in electronic business and is a junior researcher at the Research Center Steyr, School of Management, in the area of digital economy. She has worked on several small projects in the field of marketing and electronic business and is currently working on Opinion Mining and Web 2.0 methods within the TSCHECHOW project.