System Learning of User Interactions

Michael Kaschesky
Bern University of Applied Sciences
ksm1@bfh.ch

Guillaume Bouchard
Xerox Research Center Europe
guillaume.bouchard@xerox.com

Stephane Gamard
salsaDev SA
stephane.gamard@salsadev.com

Adrian Gschwend
Bern University of Applied Sciences
gsa1@bfh.ch

Patrick Furrer
EuResearch
patrick.furrer@euresearch.ch

Reinhard Riedl
Bern University of Applied Sciences
rer2@bfh.ch

ABSTRACT (REQUIRED)

The case presented in this paper describes an early prototype and next steps for developing a user-adaptive recommender system using semantic analysis and matching of user profiles and content. Machine learning methods optimize semantic analysis and matching based on implicit and explicit feedback of users. The constant interaction with users provides a valuable data source that is used to improve human-computer interaction and for adapting to specific user preferences. This can lead to, among others, higher accuracy and relevance in content matching, more intuitive graphical user interfaces, improved system performance, and better prioritization of tasks.

Keywords (Required)

machine learning, human computer interaction, crowdsourcing, knowledge management.

INTRODUCTION

Due to the number of different themes and types of funding channels, research funding programs are becoming increasingly complex to be exploited by smaller institutes and companies. The case presented in this paper describes an early prototype and next steps that offer a viable solution to the problem based on semantic analysis and matching of profiles to research themes and calls. Machine learning methods optimize semantic analysis and matching based on implicit and explicit feedback of users [7]. The constant interaction with users provides a valuable data source that is used to improve human-computer interaction and for adapting to specific user preferences. This can lead to, among others, more intuitive graphical user interfaces, improved system performance, and better prioritization of tasks.

BACKGROUND AND PROBLEM

Perhaps nowhere else than in ground-breaking research and innovation is the need for large-scale cooperation as clear. The fragmented landscape of national research programs in Europe hindered the type of cooperation needed for technology leadership. In 1984, this led to the creation of pan-European research programs, the so-called Framework Programs for Research and Innovation. Within 30 years, the funding increased from 3.75 billion Euros for a four-year to an estimated 80 billion Euros for a six-year period starting 2014. Correspondingly, not only the number of themes and calls increased but also the types of funding channels in the Framework Programs. In addition, national research programs also expanded and so did their themes and calls.

National Contact Points (NCPs) for research areas were established in each participating country as the main provider of advice and individual assistance for researchers and organizations interested in research funding opportunities. However, research institutes and innovative companies who may benefit from funding programs may never find a funding opportunity that matches their competences and work. The Swiss NCP for ICT summarizes the problem:

"SMEs interested in public funding programs have to find the right sub-program, the right call, the right topic, and the right partners. This is a very challenging task and can be very complex, because in addition to the thematic opportunities there are different horizontal programs such as SME-specific measures. An SME would have to scan the entire database or subscribe to broad areas of interests. This is not a very user friendly solution and is not in favor of having more SME participation."

Due to this lack of transparence, a whole industry sprung up whose main business is to consult research institutes and small companies in finding the right funding opportunity.

VISION AND SOLUTION

Together with an innovative start-up in text mining and semantic matching, the Swiss network of NCPs Euresearch developed the vision of matching profiles of research institutes and innovative companies to the themes and calls specified in the various funding programs. The first prototype was presented to a small audience and received strong support from national and international experts and representatives. It features a semantic-aware search engine that goes beyond full-text search and keyword-to-keyword matching:

- Fuzzy keywords extend the keyword search query entered by the users with synonyms, related words, and linked data sourced from existing dictionaries and semantic-web data.

- Contextual search provides an in-depth analysis and understanding of each word within its context and best possible matches. Users simply paste or drag and drop a sentence, paragraph, or even a document and key concepts and a list of relevant results are displayed.

- Related documents when browsing through funding opportunities display at a comprehensive, highly relevant list of all contextually related opportunities and possible project partners.

- Automatic organization of data by categories and keywords. The user can navigate through the information leveraging the structured data.

Researchers from the regional university of applied science and from a worldwide leader in document systems and services joined the effort in order integrate machine learning algorithms for user adaptation, crowdsourcing, and system suggestions. Machine learning based on user interactions is used not only to optimize the matching accuracy but also to the way results are visually displayed in a user-preferred way.

The core of the innovation concerns the target group-specific optimization of matching and result display based on transfer learning from individual users. Instead of optimizing results only individually per user, the system learns from the anonymized user interactions to derive optimizations for specific groups of users.

INFORMATION MINING AND INTERLINKING

The text mining process consists of text feature extraction, named entity recognition, coreference resolution and entity normalization. Named entities (e.g. names of people, organizations, locations, addresses) can be regarded as special tokens which often span several domains but act as single semantic units. Extracted text features are put in context to provide unambiguous results. The resultant terms and knowledge is interlinked with information available in the Internet/Semantic Web, such as Linked Open Data (LOD).

The power of the semantic web lies in two simple things: A simple, scalable data model combined with links (or URIs, Uniform Resource Identifiers) to uniquely describe and identify knowledge. The data model is using a simple subject-predicate-object combination to form so-called triples. RDF is the propagated data model for triples in the semantic web and makes it possible to store this knowledge in both human and machine usable form. The fact that machines can access this knowledge makes it possible to perform tasks on huge amounts of data which otherwise would have to be carried out by human experts.

The real power of LOD lies in interlinking data sets with each other. DBPedia for example exposes all facts that can be exported from a Wikipedia entry in RDF, while ProductDB gathers data about any kind of product. ProductDB links to DBPedia whenever possible, as a result we can derive information about the same thing from two different data sources. For that, a unique URI describing a thing exists in both data sources, which in itself is the interlink between the two data sources.

KNOWLEDGE FINDING AND MATCHING

The intelligent search described above provides a quick and efficient way for users to search, refine and filter relevant information. In addition, and to enable navigating through the results, semantic matching extracts contextual relationships between all content indexed by the system. As such, the matching can provide a list of suggested documents and articles based on the content the user is currently viewing. This list of suggestions can be organized by categories to further enhance its relevancy and positive impact on the user's search experience. An experimental feature involves exposing facets within related content leading to something like "search guided navigation" of content.

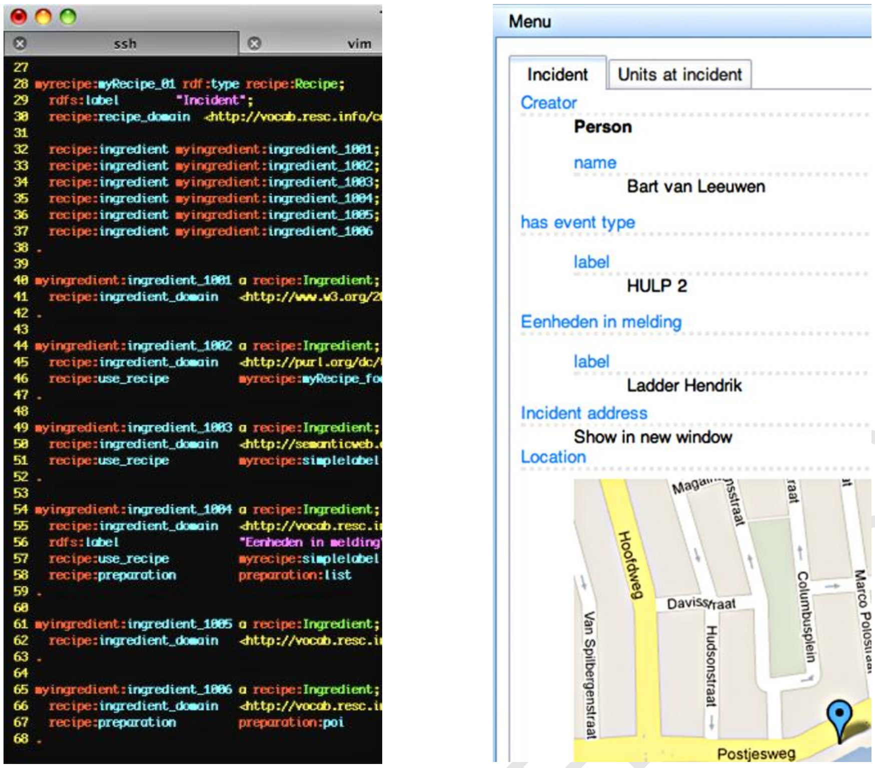
Adaptive search relates to the fact that search results improve and are aligned to user preferences based on the analysis of user implicit and explicit feedback. Adaptive search techniques are related to the learning to rank paradigm [3]. Recent advances in customized search are based on multi-task ranking techniques which enable a good trade-off between a user-independent search engines (high coverage but low precision) and fully customized systems (every user has different search result, leading to a small coverage but a high precision of the results).

A technique called query intent discovery refers to structuring and interlinking an unstructured query text submitted by the user in order to improve ranking and relevance of results. It is related to the relevance feedback principle where the search interaction is composed by multiple query refinement steps. Query intent discovery is accomplished by recognizing entities in a user query and thereby deriving a machine-understandable query statement ([3], part 1.2).

INTELLIGENT AND DYNAMIC DASHBOARDS

The user dashboard is "intelligent" meaning that its layouts and designs are adaptive based on device characteristics and user profiling and customizations. It is coded in one environment with all functionalities accessible from both desktop PCs as well as mobile devices (e.g. laptop, tablet, smartphone) where the layout and design (e.g. element position, screen size) automatically adapt to the device characteristics and user customizations. The dashboard is also highly modular where the major functions are used individually based on user profiling and customizations.

The key to flexible, modular visualization is to take advantage of the most important feature of semantically stored data, namely being self-descriptive and machine "understandable". This semantic framework enables the system to generate dynamic user interface for any given situation on the fly, independent from a given process or task. This is done by identifying the type of the data entities to be represented, and then selecting the most appropriate representation for it, e.g. a string for a name or a bitmap for a street map. The overall layout of the representation is dynamically determined by the structure of the data itself, which can, where required or intended, be overridden by additional data from the system or the user. Active learning techniques based on the value of information are used to decide which refinements should be proposed to the user [5][6].



```
myrecipe:myRecipe_01 rdf:type recipe:Recipe;
  rdfs:label        "Incident";
  recipe:recipe_domain  <http://vocab.resc.info/c

  recipe:ingredient myingredient:ingredient_1001;
  recipe:ingredient myingredient:ingredient_1002;
  recipe:ingredient myingredient:ingredient_1003;
  recipe:ingredient myingredient:ingredient_1004;
  recipe:ingredient myingredient:ingredient_1005;
  recipe:ingredient myingredient:ingredient_1006
.

myingredient:ingredient_1001 a recipe:Ingredient;
  recipe:ingredient_domain   <http://www.w3.org/2
.

myingredient:ingredient_1002 a recipe:Ingredient;
  recipe:ingredient_domain   <http://purl.org/dc/
  recipe:use_recipe          myrecipe:myRecipe_fo
.

myingredient:ingredient_1003 a recipe:Ingredient;
  recipe:ingredient_domain   <http://semanticweb.
  recipe:use_recipe          myrecipe:simplelabel
.

myingredient:ingredient_1004 a recipe:Ingredient;
  recipe:ingredient_domain   <http://vocab.resc.i
  rdfs:label                 "Eenheden in melding'
  recipe:use_recipe          myrecipe:simplelabel
  recipe:preparation         preparation:list
.

myingredient:ingredient_1005 a recipe:Ingredient;
  recipe:ingredient_domain   <http://vocab.resc.i
.

myingredient:ingredient_1006 a recipe:Ingredient;
  recipe:ingredient_domain   <http://vocab.resc.i
  recipe:preparation         preparation:poi
.
```

Creator
Person
name
Bart van Leeuwen
has event type
label
HULP 2
Eenheden in melding
label
Ladder Hendrik
Incident address
Show in new window
Location

Figure 1: The retrieval of required data items from diverse sources is coded dynamically (left) depending on the information needs and composed into a situation-aware layout (right)

The flexibility of the framework enables users to modify how data is presented to them in a given situation. This is achieved by storing a personal, modified set of layout data for a given result set and situation per user. This feature greatly benefits from the generic property of self-described data, as it does not have to be designed, implemented and maintained per use case. The framework is also used to easily bind external web services to the solution, providing additional means of interaction with displayed data. Having once identified the types of required input and the expected output values of a given web service, it can be used to dynamically provide data for a given result set and modify it as if it was an internal service.

KEY INNOVATIONS OF THE SOLUTION

The key innovations are related to the development of machine learning algorithms covering the following T areas:

- User adaptation – components of the system are not static but evolve over time by taking into account user implicit and explicit feedback;
- Crowdsourcing – users/administrators must not spend large amounts of time curating content and giving explicit feedback, they focus on curating only the content involved to complete their task;
- System suggestions – customization means that the system actively but unobtrusively interacts with the user in order to improve its performance as experienced by the user.

The constant interaction with human users provides itself a source of information that is used to improve human-computer interaction and for adapting to the specific user needs and preferences. This can lead to, among others, more usable graphical interfaces, improved system performance, and better prioritization of tasks. Crowdsourcing refers to methods for tapping into the collective intelligence of participating users to more efficiently complete tasks traditionally done by a single person or small group. System suggestions concern pro-active but unobtrusive questions by the system based on learned user behavior aimed at introducing features the user may like.

The following table summarizes the different components and techniques of user adaptation, crowdsourcing, and system suggestions:

Challenges / Methods	User adaptation	Crowdsourcing	System suggestions
Information mining and interlinking	**Advanced data curation (supervised classification)**	**Crowdsourcing for advanced data curation (multi-task learning)**	**Automating advanced data curation (active learning)**
Knowledge finding	**Adaptive searching (learning to rank)**	**Customized searching (multi-task ranking)**	**Search refinement (relevance feedback)**
Knowledge matching	**Semantic matching (relational model learning)**	**Crowdsourcing for semantic matching (tensor factorization)**	**Automating semantic matching (active relational learning)**
Graphical user interfaces	**Adaptive layouts (customized views)**	**Shared visualizations (transfer learning)**	**Feature suggestions(preference elicitation)**

Table 1: User adaptation, crowdsourcing, and system suggestions applied to the main challenges of the project

EXAMPLE ILLUSTRATION AND MATCHING METHODOLOGY

The underlying methods for matching content are based on text similarity matching algorithms developed by salsaDev. The salsaDev matching service is context sensitive. That is, each document has a set of parameters attached to it and it is possible to apply a selection of the documents satisfying the constraint before running the actual matching algorithms. For example, one parameter could be the EEN type (technology offer, request, business offer, request or partner search). A constraint could limit the search of opportunities to the technology offers. The search would be then restricted to these opportunities. Constraints are expressed as key-value pairs and provide the functionality to execute the search within a sub-set of the index.

However, it is not within the scope of this paper to describe the underlying methods, rather, to describe optimization methods, especially using user profiling, that provide the functionality needed for the end user, in this case a regional development agency.

Euresearch is currently facing the following difficulties:

- Euresearch supports its clients not only for FP7 or EEN but also for many other programs (COST, ERA-Net, IMIs, etc.). Each program comes with its own specific classification structure (Health, ICT etc.) and makes it impossible to manually profile the client in all these schemes.
- The past experience shows that it is difficult to motivate the client to profile her/himself over keywords. The opportunity finder provides a simple way to gather clients interest by investigating the selected opportunities. The use of supplementary information (selected opportunities, free text) for the profile and its benefit has to be investigated.
- The client receives currently e-mails based on his/her FP7 profile (keyword based). The existing FP7 keyword list has to be assessed and possibly modified and improved.

Optimized content matching using user profiling techniques

The user profiling makes extensive use of search and categorization. This section gives a short description of the basic ideas behind the profiling.

In information retrieval the task is, given a query q and a set of documents (or opportunities in our case) {d1,d2,...,dN}, to rank the documents according to their similarity to q, the highest rank providing the most relevant documents. If a user profile can be treated as a query, promoting content based on a user-profile is equivalent to find the documents which best matches the profile.

Unfortunately, most of the client's profiles are not expressed in terms of a query q, but in terms of categories. Typically a person’s interests are expressed as a set of categories {e.g. “Safety”, “Environment”, “Waste Management”}. The categorization/ classification problem aims at assigning a record (document or opportunity or person's profile etc.) to one or more categories within a predefined set of categories {c1,c2,...,cM}.

In our case the profile is the combination of up to three elements:

- Categories: a set of categories of interest to the user. These categories are expressed in a given taxonomy.
- Opportunities: a collection of opportunities of interest to the user (similar to bookmarking of opportunities).
- Free text: a raw text query (keyword or sense) describing the user's interest.

These three elements will be combined as illustrated in the table below. Each diagram represents a pseudo “document-index” space. The document index or categorization space is generally a high dimensional space (50 to 1500 dimensions) but is illustrated here in a 2-dimensional space with terms 1 and 2.

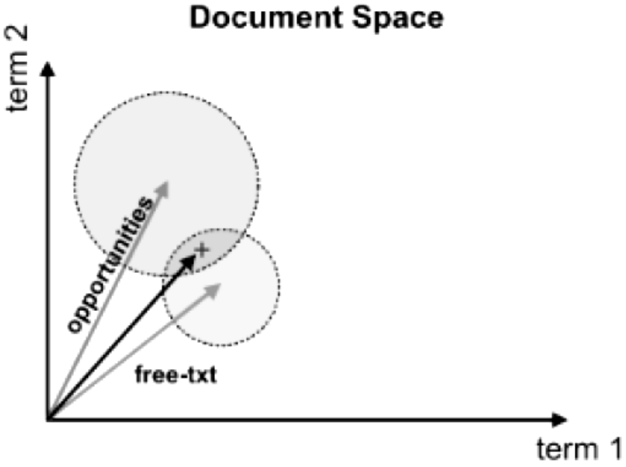


Figure 2: Components in the document-index space with lower vector computed from user profile and upper vector based on opportunities where overlap defines the region containing opportunities of interest

SalsaDev provides a search service capable of querying its index based on either raw text (used for user-profile query) or a set of opportunities (pre-indexed content - currently used in the “similar” content widgets of the Opportunity-Finder). Figure 1 above shows both profile components in the document-index space: the lower vector is computed from the free-text of the profile, the upper vector is given by the list of bookmarked opportunities and the overlap of both defines the region containing opportunities of interest for the respective components of the profile. While it is impossible to merge both types of queries both result sets are comparable and space equivalent, i.e. they can be merged at no additional computational costs.

Semantic Similarity Matching

There are 3 major steps to provide results based on a user profile.

- Information Access (search): By executing a search based on the user's bookmark and its query we are creating a first sub-set of the index which is most relevant to the user's interest. It is important to keep in mind that this subset can be the whole index itself (although the result set will be ranked accordingly to query and bookmarks). This information access step (which is 2 merged searches) can also be used to constrain the set from which the recommendations will be made (constrains).

- Aggregation (merge): this step is more an incremental set, preparing for the final step. During merger heuristics drive the weighting of independent searches. Weights can be applied for constrains (are news more important than blogs for example), for metadata (i.e. latest news rather than old ones) and sense-heuristic (is the query a set of keywords or a complete CV?)
- Evaluation: The final set, and most computationally expensive, of the recommendation process. The evaluation step transforms the result set generated by the information access process by ranking its aggregated results with respect to the categories of the profile. There are a couple of key points such as a) evaluation and information access are completely independent with respect to the taxonomy; b) not the raw score of categorization ranks the result, but the computed intersection (relevancy in each process) of the fit; c) even if information access fails, evaluation will sort the constrained set accordingly to categorization scores.

The first example concerns the semantic search which utilizes the text similarity matching algorithms in order to compare and retrieve text with the similar meaning. In this case, the user inputs a paragraph or more text, for example:

« While flexible pressure and temperature sensors are relatively easy to realize, gas sensors are more challenging due to their more complex structure involving gas sensitive materials that must have access to the external environment. Particularly for gas sensors targeting smart textile applications, the challenge is significantly greater since the industrial weaving is a very abrasive process. The main target of TWIGS is to design, fabricate and characterize a mechanically robust thin-film capacitive gas sensor platform compatible with industrial weaving processes. The targeted sensor platform is particularly interesting to develop textile integrated sensors arrays with each individual sensor functionalized with a different sensing material to perform complex sensing tasks »

The list of results is retrieved as described above. By selecting the first opportunity (e.g. clicking on the title) one reaches the detailed description of the opportunity (in that case a project description of an FP7 project under the Food, Agriculture and Biotechnology theme). On side panels, additional opportunities that are closely related (semantically speaking) with that opportunity are presented, to allow further exploration. Scientific officers at Euresearch confirm that they would have advised a client to look into this same opportunity, given the original query the client had formulated.

It is also possible to use a few keywords as in most internet search engines to specify the query, such as “smart textiles“.

In this case, the Opportunity Finder proposes a technology opportunity in «conductive ink for smart textiles», a business opportunity in «trade intermediary services for smart casual wear» and one technology request in «cosmetotextile lotion».

After selecting via filtering only those R&D sub-programs in nanotechnology, the list of opportunities is completely changed and focuses now specifically on opportunities related to the NMP program (where it really belongs to), and further down in the list to the ICT program (smart components and integrated systems).

The remaining sections describe the procedural steps, not the technical details, to accomplish the opportunity matching. The procedural steps serve to illustrate and explain the overall approach. It is, however, not within the scope of this paper to detail the technical specifications, methods, and algorithms underlying each of the steps.

CONCLUSION

This paper presented the case of an early prototype for semantic analysis and matching of user profiles to funding opportunities of research programs. The novelty of the described system concerns the utilization of data from implicit and explicit user interactions for improving the accuracy of matching and the user-specific layout of the dashboard elements and contents. It is shown how data from user interactions are used to improve human-computer interaction and for adapting to specific user preferences.

However, the case study presented in this paper is based on a prototype and testing among small user groups. Large-scale empirical evidence of the effectiveness of the described machine learning methods for optimizing semantic matching based on user feedback must yet be collected.

The next steps involve the tighter integration of LOD to enable the system to recognize the specific taxonomies employed by users and to learn from multiple user contributions. The multi-task learning algorithms enable the system to classify new text items automatically based on existing classifications of similar texts. These machine learning and crowdsourcing methods for taxonomy development and maintenance minimize manual intervention of users and maintainers in data curation.

Ontologies and their taxonomies contain several thousands of entries, so automated tools to support the creation of new annotations and efficiently annotating content are needed. Supervised classification naturally fits this task. The user annotates content with own annotations and once the classifier learned how content and annotations are matched, the system suggests to the user similar content that is automatically annotated with similar annotations. If the user confirms, the system continues to annotate the type of content with that set of annotations. The more confirmations the system receives, the higher the accuracy of the resulting annotations [4].

Concerning the further directions, the approaches and innovations used for the described recommendation system for a research funding programs can be extended to other uses of recommendation systems. For example, one extension to this approach is planned to aid first responders in emergency services who are increasingly confronted with a flood of separate information and devices that are hardly manageable by one person. The goal is to combine relevant data from several sources, such as object descriptions, hazard location plans, and data on wind direction in order to recommend predefined actions to the appropriate first responders, for example, based on context of the situation and their role and location. Together with a broad network of project stakeholders, several prototypes and an infrastructure is planned for the timely, context-aware, and role- and needs-based delivery of relevant emergency information.

REFERENCES

[1] Belew, R.K.. Finding Out About, Cambridge Univ. Press, 2000.

[2] [3] Cheung, J. C. K. and Li, X. Sequence clustering and labeling for unsupervised query intent discovery. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12). ACM, New York, NY, USA, 383-392, 2012.

[3] Joachims, T., F. Radlinski. Query Chains: Learning to Rank from Implicit Feedback. Proc. of ACM SIGKDD, 2005.

[4] Maynard, D., A. Funk & W. Peters. NLP-based support for ontology lifecycle development. Proc. of ISWC Workshop on Collaborative Construction, Management and Linking of Ontologies, 2009.

[5] Reichart, R., K. Tomanek, U. Hahn, A. Rappoport. Multi-task active learning for linguistic annotations, Proceedings of ACL, 2008.

[6] Settles, B. Active Learning Literature Survey. Computer Sciences Technical Report 1648. Univ. of Wisconsin, 2009.

[7] Stumpf, S. et al. Toward Harnessing User Feedback for Machine Learning. Proc. IUI, 2007.