Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 104 (2017) 539 – 547

Procedia Computer Science

ICTE 2016, December 2016, Riga, Latvia

Search Intention Analysis for Task- and User-Centered Visualization in Big Data Applications

Dirk Burkhardt[a,b,*], Sachin Pattan[a], Kawa Nazemi[a,b], Arjan Kuijper[b]

[a]*Fraunhofer-Institute for Computer Graphics Research (IGD), Fraunhoferstr. 5, 64283, Darmstadt, Germany*
[b]*TU Darmstadt, Fraunhoferstr. 5, 64283, Darmstadt, Germany*

Abstract

A new approach for classifying users' search intentions is described in this paper. The approach uses the parameters: word frequency, query length and entity matching for distinguishing the user's query into exploratory, targeted and analysis search. The approach focuses mainly on word frequency analysis, where different sources for word frequency data are considered such as the Wortschatz frequency service by the University of Leipzig and the Microsoft Ngram service (now part of the Microsoft Cognitive Services). The model is evaluated with the help of a survey tool and few machine learning techniques. The survey was conducted with more than one hundred users and on evaluating the model with the collected data, the results are satisfactory. In big data applications the search intention analysis can be used to identify the purpose of a performed search, to provide an optimal initially set of visualizations that respects the intended task of the user to work with the result data.

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016

Keywords: User-centered-design; Human-computer-interaction; Information retrieval; Predictive analysis; Interactive search

1. Introduction

Nowadays search engines are designed to search for clear objectives, such as concrete resources, products or information, which results are almost presented in list representation ordered by complex algorithms that majorly focus on keyword matching. Advanced search systems that use even graphical visualizations are able to visualize

* Corresponding author. Tel.: +49-6151-155-578; fax: +49-6151-155-139.
E-mail address: dirk.burkhardt@igd.fraunhofer.de

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016
doi:10.1016/j.procs.2017.01.170

results beyond result listings. These advanced visualization opportunities allow showing structures or in general specific aspects such as geographical or temporal properties visually. In fact that means, searches are not limited on clear target searches and therewith objectives. Modern information retrieval systems can cover also exploratory searches, e.g. to gather an overview about unknown topics or complex linking of entities, resources and objects. But also analysis is a typical tasks for modern information retrieval systems, which aim lays on analyzing certain properties or complex numeric values.

The challenge of modern information retrieval system is to show the results with appropriate visualizations or visualization algorithms. If users perform searches about persons, e.g. the biography, a simple result listing ordered by keyword matching is sufficient. But what if users aim to compare different cameras by their properties to find the best option to buy? Or what if users aim to search for German renowned search institutions that focus on renewable energy? The list presentation is not that helpful to start the data analysis or exploration.

To be able to show the results in a beneficial set of visualizations, to know the search intention would help to better identify what visualization might be useful. Indeed, the idea of identifying the search intention is not new. The current existing approaches can be classified in semantically analysis[1] and generic/situational query analysis[2,3,4]. The semantical query analysis approaches is language depending and requires predefined rules. The generic or situational query analysis using only basic information, such as query length, or additional behavioral information like the geographical location or details about the used (mobile) device. However, all of the approaches are not generic and therefore have deficits in particular in normal search contexts.

Therefore, this papers describe a novel approach to identify the intention of a performed search just on the basis of the entered search query, and majorly on the basis of the word frequency that are used in the search phrase.

2. Search intention analysis approach

In this section we describe the principle approach to calculate the search intention majorly based on the word frequency. Before the method can be described, we explain the preliminary study we performed before, to identify relevant influencing factors to calculate the user's search intention.

2.1. Preliminary study

Since there is no existing work to identify search task majorly based on word frequency, we had to perform a preliminary study to identify rules or patterns. Therefore we asked more than 50 people, almost with a computer science background, to check their search history on Bing, google etc. to send us real performed search phrases. Together with the search phrases, we ask them to classify the searches, if they would identify them more likely as target, exploratory or analysis search. Even though the people could provide some notes, e.g. if they were not sure what the goal was most likely.

On the basis of this data we tried to identify rules or patters on which basis we could classify a search phrase. Even more we developed a metric to be able to calculate the search intention into the task categories: targeted, exploratory and analysis search.

2.2. Rules and patterns to identify search intentions

With this classification information in hand, we tried to recognize some patterns, which would help us in forming the hypotheses. Based on these found hypotheses, six were considered for classifying a query as exploratory, targeted or analytical search query. As there were only few analytical search queries, only few patterns/hypotheses were recognized for them (see Fig. 1):

- Query length for Exploratory: If the sentence (query) contains less or equal than two words, it is most likely to be an *exploratory search* query
- Most frequent Word in the query: If the most frequent word in the query has a class less than five, then the query is most likely to be *targeted search*

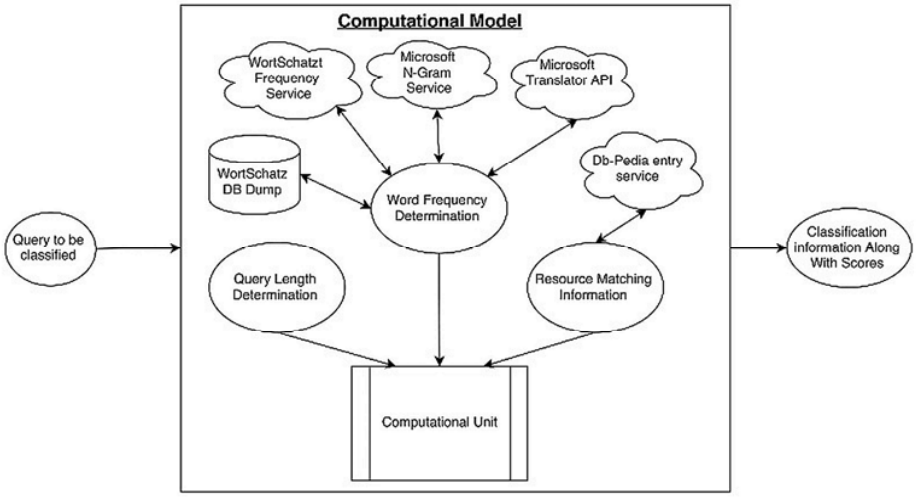


Fig. 1. Abstract computation model of the search intention analysis with all the included components.

- Presence of words such as Year, Age, Birth date etc.: Presence of such words in the query seemed to indicate that the query would be *targeted search*. But, as this would be on the basis of linguistic/semantic analysis and therefore language-dependent analysis (there can be many such cases as per the grammar), this hypothesis was dropped
- Query Length for Analytic queries: As per observation from the survey data, the queries with length greater than six tend to be *analytical search* queries
- The highest frequency word: In the query with length greater than six, the *analysis search* queries have at least one word in the last four words of the query which falls in the frequency class greater than ten
- Existence of a Number: In the query with length greater than six, the *analysis search* queries have at least one word which is a number in the last four words of the query
- DBpedia Lookup entry: If there is a direct hit (an entity could be found and has a class reference) of the query specified in DBpedia and there is an existing entity, then it must be *targeted search*. As the user is looking for something specific. Or else, the query will be trimmed word by word from the right to calculate how much the query is more likely a *targeted search*
- DBpedia Lookup entry for whole query but not a direct hit: If there is a hit for the whole query in DBpedia, but there is no class reference, then it is more likely to an *exploratory search*. As the query represents a generic entity in the web search and possibly a category, bullet word 30 concept etc. that summarizes multiple entities under and therefore an overview about the relating entities is beneficial. So, the user wants to explore about it than he already knows about it.

Based on the identified patters we could define a computational model that checks a given query against the defined rules and in consequence calculates the score for each kind of search task (see **Error! Reference source not found.**).

2.3. Search intention calculation

The search intention is identified based on the highest score for the given search task, in particular the *Exploratory Score* (*ES*), *Targeted Score* (*TS*) and *Analytical Score* (*AS*). Therefore the scores are calculated in perspective of the above identified patterns. The calculation is based on the *Query Length Factor* (*QLF*), *Least Frequent Factor* (*LFF*), *DBpedia Information Factor* (*DIF*), *Analytical Factor* (*AF*), *DBpedia Score* (*SD*) and the *Query Length* (*QL*). We could identify the following score calculations for the mentioned search tasks:

$$ES = \sqrt{(QLF)^2 + (1 - LFF)^2 + (1 - DIF)^2 + (1 - AF)^2} \tag{1}$$

$$TS = \sqrt{(1 - QLF)^2 + (LFF)^2 + (DIF)^2 + (1 - AF)^2} \tag{2}$$

$$AS = \sqrt{(1 - QLF)^2 + (LFF)^2 + (DIF)^2 + (AF)^2} \tag{3}$$

The concrete attributes are calculated as follow:

$$QLF = \frac{2}{n} \qquad \text{n…number of terms in the query} \tag{4}$$

$$LFF = \frac{(10 - c)}{5} \qquad \text{c…class of least frequent word} \tag{5}$$

$$DS = \begin{cases} 2 \cdot \dfrac{q}{QL} & \text{If whole or part of query matches to a dbpedia resource and has class references} \\ & q\text{... length of the matching query part } (q \le QL) \\ 0.5 & \text{If whole or part of query matches to a dbpedia resource, but has no class reference} \\ 0 & \text{Else (in particular if no match to a dbpedia resource)} \end{cases} \tag{6}$$

$$DIF = \max(DS) \tag{7}$$

$$AF = \begin{cases} AF + 0.3 & \text{If } QL > 6 \\ AF + 0.2 & \text{for each non-consecutive occurrence of a Number in the query} \\ AF + 0.1 & \text{If } QL > 6 \text{ and a word with frequency Class > 10 exists in last four words of a query} \\ 0 & \text{Else} \end{cases} \tag{8}$$

The mentioned formulas are the most relevant for the calculation. Some of the formulas refer to a so called frequency class, which is introduced in the next section. In general the frequency classes are simplified a method to categorize the frequency ranges into a fix number of categories. This makes it easier and better comprehensible than using real frequency values.

2.4. The role of word frequency

This component depends on language of the query. Different languages require different sources of information as there are language specific sources of the frequency information. For instance, there are several corpora for English word frequencies while there are only few such corpora for German word frequencies. Leipzig Corpora Collection (WortSchatz)[5,6], which has the representation for corpora in different languages using the same format and comparable sources was used to get the frequency information. WortSchatz provides a SOAP (Simple Object

Access Protocol)[7] service for German word frequency information whereas for other languages, they provide database dumps corresponding to each language. WortSchatz was used for English and German and the model can be extended to support further languages by downloading the corpora (database dump) from WortSchatz. The frequency class determination can be done as:

$$N = \left\lfloor \log_2 \left(\frac{\text{Frequency of the most frequent word}}{\text{Frequency of this word}} \right) \right\rfloor \tag{9}$$

Also, the Microsoft Web N-gram Services[8] was used to get the frequency information for almost all the languages. As the intention is to get the frequency of individual words and not the whole query or part of the query, the 1-gram service serves the purpose. In the 1-gram service, the frequency of words are given by the log likelihood of a word.

That is how likely a given word is used. It will usually be a negative value. The lower the likelihood value, the more frequent is the word used. For instance, the most frequent word in English 'the' has the Ngram log likelihood value of -1.4985, whereas the rarely used word 'hock' has the likelihood value of -6.303. Based on the Ngram values, we categorized the words into frequency classes by applying the base-2 algorithm[9] on the anti-log values of their log likelihood values. The class *N* can be given as:

$$N = \left\lfloor \log_2 \left(\frac{\text{antilog}\left(\text{Log Likelihood of most frequent word}\right)}{\text{antilog (Log Likelihood of this word)}} \right) \right\rfloor \tag{10}$$

In this classification, the most frequently used item (e.g. 'the') belongs to frequency class 0 (zero) and any item that is approximately half frequently used that of 'the' belongs to class 1. In the example given before, the word 'hock' with likelihood value of -6.303 belongs to class 16.

The class distribution does not compel to the language independent feature of the model. As the Ngram values are not language specific, it was difficult to determine the frequency classes for words of languages other than English as the Microsoft Ngram service gives the log likelihood, which is the absolute probability of the occurrence of the word than the probability with respect to the most frequent word of the language to which the given word belongs to. As a result, it may happen that the least frequent English word and the most frequent word of another language might end up having the nearby Ngram values. As a result, both end up having the same frequency class values.

To face this problem, the concept of calculating an 'offset' was used. This offset is added to all the Ngram values of the Non-English words. The offset for a given word is basically the difference in the Ngram value of a moderately used word in English (for e.g. 'dog') and in the language to which the given word belongs to. For instance, if the word is 'der' (one of the most frequent words in German) for which the frequency class is needed, the offset and the reformulated Ngram value can be given as below:

$$Offset = \text{NValueRefWordEnglish} - \text{NValueRefWordLanguage} \tag{11}$$

$$\text{NValueWordLanguage} = \text{NValueWord} + \text{Offset} \tag{12}$$

The offset = Ngram value for 'dog' in English (-4.0375) - Ngram value for 'dog' in German ('Hund') (-6.685) =2.6475. So, the Ngram value for the word 'der' = Original Ngram value (-4.444) + offset (+2.6475) = -1.7965, which assigns it to class 0 while it was assigned to class 10 before.

The described approach is used for all other languages and word of those languages too. The information from both sources was combined for English and German queries where as for other languages, only N-gram was used.

The Frequency information was represented in terms of the frequency classes in which, the lower the class the more frequent is the word used in particular language. For instance, if the word has a frequency class 0, it is one of the most frequent words in the language. And, a word with frequency class 5 is not so frequent compared to a word with frequency class 0.

In general, if the word's frequency class is less than five, it is considered as a more frequent word, and if not as less frequent word. For instance, as per hypothesis two, if the query has a word whose frequency class is less than five, it is more likely that the query falls into the 'targeted search' category as stated in the equations 1 to 3. Also, as per the hypothesis five, in the same set of formulas we can see that if the query whose length is greater than six and if the query has at least one word whose frequency class is greater than ten, it is more likely that the query falls into the category of 'analysis search queries'.

2.5. Distance supervision with DBpedia

If there is an existing entity with the same label (contents of label) in an online resource repository such as DBpedia[10] and the resource has a class definition, it can only be a 'targeted search' as the query represents the most known entity in Web search and the user does not intend to explore about it. So, this information had to be given higher weight than other factors. The more the part of a query matches the exact entity name of the resource which has a class definition, the more is the query a 'targeted search'.

However, when there is a hit for the whole query, but the resource does not have a class definition, the query is more likely to be an 'exploratory search'. This is because the searched query is a more generic term, which does not fall under the category of the persons, places, books etc. This indicates that the user intends to explore about the query an aims to retrieve an overview about the topic. The equations 1 to 3 indicate how the resource matching factor (*DbPedia Information Factor*) determines the category of a query to be a 'targeted search' query, if it is a direct hit and when there is a simple hit for the whole query, it is more likely to be an 'exploratory search'.

2.6. Implementation: A Web-service for search intention analysis

The implementation is made as a simple REST web-service, which enables an easy integration in a couple of external applications. The request needs to have the format *http://server.tld/WordFrequencyService/rest/frequency/[search phrase]* and the response will be given in XML as shown in Fig. 2. Next to the most important calculation of the scores for explorative, targeted and analytical search, even further details e.g. about the key phrase language and also details to the containing words of the key phrase will be given.


```
▼<frequencyDetails>
   <query>mobile phone</query>
   <language>ENGLISH</language>
   <queryType>Explorative</queryType>
  ▼<calculationDetails>
     <minimumNormalizedClass>9</minimumNormalizedClass>
     <explorativeScore>1.7</explorativeScore>
     <targetedScore>1.1357816691600546</targetedScore>
     <analyticalScore>0.5385164807134505</analyticalScore>
   </calculationDetails>
  ▼<word-details>
   ▼<words>
      <text>mobile</text>
      <wordLanguage>ENGLISH</wordLanguage>
      <frequency>144</frequency>
      <wortSchatzFrequencyClass>11</wortSchatzFrequencyClass>
      <ngramValue>-3.486</ngramValue>
      <ngramClass>7</ngramClass>
      <ngramOffset>0</ngramOffset>
      <normalizedClass>9</normalizedClass>
    </words>
```

```
▼<words>
    <text>phone</text>
    <wordLanguage>ENGLISH</wordLanguage>
    <frequency>481</frequency>
    <wortSchatzFrequencyClass>10</wortSchatzFrequencyClass>
    <ngramValue>-3.514</ngramValue>
    <ngramClass>7</ngramClass>
    <ngramOffset>0</ngramOffset>
    <normalizedClass>9</normalizedClass>
  </words>
 </word-details>
</frequencyOetails>
```


Fig. 2. Result of a performed query phrase analysis for 'mobile phone'.

The inclusion in existing information search application as REST service makes it easy to advance existing application with the ability to initially show a more appropriating composition of visualization that show the search results.

3. Use-case: Support in information systems

The major intention to create this web-service is the fact that in particular graphical information systems supporting a couple different search options and even more a variety of visualization algorithms to show the results for analysis. In fact it is difficult to preselect the best visualizations for any kind of search task, since it is normally unknown what kind of search type the user has in mind. The developed systems solves this problem and reduces the general search intentions at least on the most relevant categories. In fact, visualization can be preselected that are appropriate for a given concrete search scenario.

The focus for using the developed system is for our graphical information systems in particular for the policy modeling and the trend analysis domain. The goal of policy modeling is to enable decision makers to make better decisions on the basis of enhancing data analysis by considering the intention of a search or analysis task[11,12]. As it can be imagine, the analysis of large and complex data sources for this purpose can be become challenging, since there are many options how data can be analyzed and for which purpose it will be analyzed. For instance, the analysis of statistical government data or the search and use of simulation models as well as simulators for foresight analysis is difficult and support in that direction is beneficial for users[13]. This is similar to visual trend analysis[14], where also here a couple of options do exist to analyze massive textual data with the goal of early identification of upcoming trends in form of innovation.

In any of the such analysis solutions where users have to deal with search interfaces, it is beneficial for the users to get support in getting appropriate initial results visualization that fit to the search intention of the users. This increases the user's efficiency on the one hand and decreases the frustration of users on the other hand, because of the fact that the user is not confronted with less useful compositions of visualizations.

4. Evaluation

As stated before, the model currently supports three categories of search-query tasks: "exploratory searches", "targeted searches" and "analytical searches". The evaluation was done separately for each category and it can be easily extended to accommodate new categories with few changes. The data from all the separate evaluations was gathered and analyzed to discuss the overall performance and efficiency of the model.

As the data collected from the survey are labeled data, the supervised machine learning evaluation techniques[15] of *precision*, *recall* and *accuracy* were applied. These techniques in turn depend on the True/False Positive and True/False Negative predictions[16] of the model. As these methods are highly recommended for any classification system, they suited the best for the model. The *precision* in the field of information retrieval is given as the ratio of the number of records retrieved, which are relevant to the total number of records retrieved (irrelevant and relevant)[16]. *Recall* in the field of information retrieval is given as the ratio of number of records retrieved, which are relevant to the number of relevant records in the repository[16].

In information retrieval, *precision* is considered as a result relevancy measure and *recall* as a measure of how many truly relevant results are returned. It is also believed that, with high *precision* and *recall*, the system will return many correctly labeled results[17]. With the same set of parameters, another measure *accuracy* can be derived. In the field of information retrieval *accuracy* is the ratio of true correct retrieval values (both true positives and true negatives) to the total number of cases examined[18].

The table and the bar-chart representations in Fig. 3 show the *precision*, *recall* and *accuracy* values of the model for the categories “exploratory search”, “targeted search” and “analysis search”, based on the True/False Positive and True/False Negative values. The results are obtained by providing the model with more than 100 queries from the survey that was conducted (testing data).

As we can see on the Fig. 3 on the left, the model performs pretty well for “targeted search” and ”exploratory search” queries in perspective of precision, recall and accuracy values. As there was less data for the analysis search queries for training as well as for testing, the results cannot be taken seriously for this category as the model could accommodate only few patterns (hypotheses) for analysis search queries. It is not clear how the model would perform with data sets consisting of many analysis search queries. The major concern being the separation of targeted and exploratory search queries, the results look good, however they can be further improved.

	Precision	Recall	Accuracy
Targeted Query	83.6%	86.44%	82%
Exploratory Query	80.64%	78.12%	87%
Analytical Query	75%	66.66%	95%

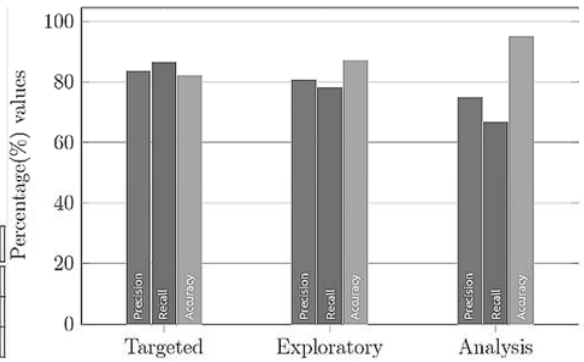


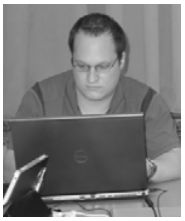
Fig. 3. The search intentions evaluation results.

5. Conclusion

The paper described a novel approach to identify the intention of a performed search just on the basis of the entered search query, and majorly on the basis of the word frequency that are used in the search phrase. It is developed as a web-service so that it can easily be used and included in modern information systems that uses in particular graphical visualizations and are therewith designed to cover a variety of search tasks, such as targeted, exploratory and analysis searches. If the web-service is included, a graphical information system can better predefine the initial visualization composition the user get as starting point for his data search and analysis. The evaluation has shown that the concept and its implementation achieved good results.

References

1. Jansen BJ, Booth DL, Spink A. Determining the user intent of web search engine queries. In: Proceedings of the 16th International *Conference on World Wide Web, WWW '07*. ACM; 2007. p. 1149-1150.
2. Mendoza M, Baeza-Yates R. A web search analysis considering the intention behind queries. In: *Web Conference 2008. LA-WEB '08*. Latin American; 2008. p. 66-74.
3. Wang Z, Zhou X, Yu Z, He Y, Zhang D. Inferring user search intention based on situation analysis of the physical world. Lecture Notes in Computer Science. Vol. 6406. Springer; 2010. p. 35-51.
4. Yi X, Raghavan H, Leggetter C. Discovering users' specific geo intention in web search. In: *Proceedings of the 18th International Conference on World Wide Web, WWW '09*. ACM; 2009. p. 481-490.
5. Biemann C, Heyer G, Quasthoff U, Richter M. The Leipzig corpora collection - monolingual corpora of standard size. In: *Proceedings of Corpus Linguistic 2007*. Birmingham. UK; 2007.
6. Quasthoff U, Richter M, Biemann C. Corpus portal for search in monolingual corpora. In: *Proceedings of the 5th international conference on Language Resources and Evaluation*. LREC. Genoa; 2006. p. 1799-1802.
7. Hirsch F, Kemp J, Ilkka J. *Mobile web services: architecture and implementation*. John Wiley. Chichester, England, Hoboken; 2006.
8. Wang K, Thrasher C, Viegas E, Li X, Hsu P. An overview of Microsoft web n-gram corpus and applications; 2010.
9. Wikipedia. Binary logarithm. Wikipedia. The free encyclopedia; 2015.
10. Lehmann J. et al. Dbpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web J*. 5; 2014. p. 1-29.
11. Burkhardt D, Nazemi K, Sonntagbauer P, Sonntagbauer S, Kohlhammer J. Interactive visualizations in the process of policy modeling. In: *Proceedings of IFIP eGov 2013*; 2013.
12. Burkhardt D, Nazemi K, Kohlhammer J. Visual Process Support to Assist Users in Policy Making. In: *Handbook of Research on Advanced ICT Integration for Governance and Policy Modeling*. IGI Global; 2014. p. 149-162.
13. Burkhardt D, Nazemi K, Ginters E, Aizstrauts A, Kohlhammer J. Explorative Visualization of Impact Analysis for Policy Modeling by Bonding Open Government and Simulation Data. In: *Human Interface and the Management of Information. Proceedings Part I: Information and Knowledge Design*. Springer. Lecture Notes in Computer Science. Vol. 9172; 2015. p. 34-45.
14. Nazemi K, Retz R, Burkhardt D, Kuijper A, Kohlhammer J, Fellner DW. Visual Trend Analysis with Digital Libraries. In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business (i-KNOW 2015)*. ACM; 2015.
15. Manning CD, Raghavan P, Schutze H. *Introduction to Information Retrieval*. Cambridge University Press; 2008.
16. Wikipedia. Precision and recall. Wikipedia. The free encyclopedia; 2015.
17. Metz CE. Basic principles of ROC analysis. *Seminars in nuclear medicine*. 8(4); 1978. p. 283-298.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12; 2011. p. 2825-2830.



Dirk Burkhardt holds a diploma in Computer Sciences from the University of Applied Sciences Zittau/Görlitz. His research interests are focused on concepts and technologies for user-centered process modeling and process support especially based on visualization technologies. He also researches in gesture-based multimodal interactions for intuitive interactions and navigations within graphical user-interfaces. Contact him at dirk.burkhardt@igd.fraunhofer.de.