World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:12, 2017

Social Media Idea Ontology: A Concept for Semantic Search of Product Ideas in Customer Knowledge through User-Centered Metrics and Natural Language Processing

Martin Häusl, Maximilian Auch, Johannes Forster, Peter Mandl, Alexander Schill

***Abstract*—**In order to survive on the market, companies must constantly develop improved and new products. These products are designed to serve the needs of their customers in the best possible way. The creation of new products is also called innovation and is primarily driven by a company's internal research and development department. However, a new approach has been taking place for some years now, involving external knowledge in the innovation process. This approach is called open innovation and identifies customer knowledge as the most important source in the innovation process.

This paper presents a concept of using social media posts as an external source to support the open innovation approach in its initial phase, the Ideation phase. For this purpose, the social media posts are semantically structured with the help of an ontology and the authors are evaluated using graph-theoretical metrics such as density. For the structuring and evaluation of relevant social media posts, we also use the findings of Natural Language Processing, e. g. Named Entity Recognition, specific dictionaries, Triple Tagger and Part-of-Speech-Tagger. The selection and evaluation of the tools used are discussed in this paper. Using our ontology and metrics to structure social media posts enables users to semantically search these posts for new product ideas and thus gain an improved insight into the external sources such as customer needs.

***Keywords*—**Idea ontology, innovation management, open innovation, semantic search.

I. Introduction

THE average useful life of a laptop is currently still two years and the useful life of high-end smartphones is scheduled by telecommunication companies with just 12 months. We have arrived at a time when customers expect shorter product and service development cycles than ever before. Companies that want to keep up with this trend have to rely on innovations and expand their innovation capability to be a core competence [1, p. 6]. According to Xu et al. [2, p. 581], the innovation process that shows a procedure model for increasing the innovation capability begins with the idea generation phase. In this phase, internal sources are used to generate ideas in the classic innovation process. Since the use of purely internal sources produced more and more products and services, which were developed beyond the customer's needs, the open innovation approach [3, p. 44] was born. External sources are used to leverage customer knowledge for the further development of products and services. According to Alt & Reinhold, the biggest source of customer-generated knowledge is in the form of social media posts [4]. To make this information usable we developed a concept to identify and structure social media posts to support the identification of relevant and potential ideas in relation to specific products. In Section II we show similar research papers and the differences to our research. Section III we give an overview of our concept. Section IV firstly gives an overview of the developed ontology and secondly gives examples of the used dictonaries. In Section V, we show in detail how we tagged the data and the tools we used to structure the social media data. Initial results and advantages of our procedure and further steps in research development are explained in Section VI Pre-Study Results and Outlook.

Martin Häusl is with the Department of Computer Science and Mathematics, Munich University of Applied Sciences, Munich, Germany (e-mail: mhaeusl@hm.edu).

Maximilian Auch is with the Department of Computer Science and Mathematics, Munich University of Applied Sciences, Munich, Germany (e-mail: maximilian.auch@hm.edu).

Johannes Forster is with the Department of Computer Science and Mathematics, Munich University of Applied Sciences, Munich, Germany (e-mail: jforster@hm.edu).

Peter Mandl is with the Department of Computer Science and Mathematics, Munich University of Applied Sciences, Munich, Germany (e-mail: mandl@hm.edu).

Alexander Schill is with the Department of Computer Science, Dresden University of Technology, Dresden, Germany (e-mail: alexander.schill@tu-dresden.de).

II. Related Work

The project *Innovation Signals*, located at the Salzburg Research Center, focuses on the early detection of weak innovation signals. Based on the theory of weak signals according to Ansoff [5] and close cooperation with partners from industry, Eckhoff et al. [6] show an approach to support the entire innovation process of companies. The researchers rely on a process with five phases consisting of methods of social market research, data analysis and consulting. The approach involves several iterations between phases one to four. The research group published [7] as a complementary analytical approach for the identification of innovation signals in online communities. The prototype is provided by means of *Apache Tomcat* [8] and uses *Apache Marmotta* [9] as well as the triplestore *Apache KiWi* [10]. The queries on the persisted data are made via SPARQL [11]. Apache Stanbol is used for content analysis. The source-specific data crawlers called *extractors* as well as transformation, sentiment analysis and search indexing are not explained in detail. Furthermore,

neither the functions nor the data presentation in the frontend nor the use of SPARQL are described in more detail. In the case of enrichment, it is merely pointed out that the extracted, ambiguous and unstructured data is linked with relevant statistics, trends and theories in order to be able to better interpret the data [6, p. 124]. However, as the manual enrichment of data from the founders of unstructuredness of data is reported and no values to metrics are specified, it is to be considered, that all analytical steps are carried out purely on the basis of the text and without measurable metrics (cf. *reputation*). The semantic processing or use of data structures such as ontologies is also not mentioned.

The approach *Mining ideas from textual information* published by Thorleuchter et al. [12] is for the identification of new and useful ideas from unstructured text. The researchers developed their own metric to measure the novelty of the idea. Patent data was used as the data source. The Basic Approach uses methods from psychology and cognitive sciences and pursues the goal of finding ideas, as it is also done by humans, in an algorithm [13].

Westerski et al. [14] also recommend the inclusion of external sources to support innovation management in companies. Westerski et al. focus on data from so-called ideation platforms such as Dell IdeaStorm [15], [14]. On such platforms, customers of a company can bring in their own ideas or improvements. Other users can then evaluate these suggestions and ideas. The form of the idea description is also done by free text. In contrast to public social media posts, however, the user is clearly encouraged to formulate an idea for a specific product. Westerski et al. propose an annotation based on a domain independent taxonomy, which is divided into the three main classes and the 11 subclasses [14]. The first step was to annotate the data by hand. Ten people annotated the same 10 ideas. In the further course of the research, an automatic annotation was carried out based on the results of the manual annotation. In this context, Westerki et al. found that automatic annotation is not appropriate for all subclasses [14].

None of the related works presented here focuses on the development of social media sources as an external source for searching for ideas on specific products. As a result, metrics that we created based on the author of a social media post have not yet been used. This gap closes our approach, which we will explain in more detail below.

III. Overview of the Concept

The main process begins by collecting social media posts from Web 2.0. Afterwards, user-centered metrics are calculated in order to classify the quality of the author and thus that of the post. In the next of the process step, the text of the social media post is semantically formatted by annotating the text. Different taggers are used for this purpose, which are explained, discussed and evaluated in more detail in Section V. Subsequently, an RDF model is generated using the ontology presented in Section III and handed over to Apache Jena Fuseki for persistence in the TDB Triple Store.

The client implemented in Angular2 accesses the data by sending corresponding SPARQL requests to the http interface offered by Apache Jena Fuseki. An overview of the entire process is shown in Fig. 1.

IV. The Idea Ontology

This section introduces the Social Media Idea Ontology (SMIO) ontology concept. At this stage of our work, the ontology was completely redesigned from scratch, as other available ontologies were either too extensive or unsuitable for the intended purpose. For modeling, W3C standards such as RDF, RDFS and namespaces were used.

1) Classes and Dictionaries: The *smio:User* class describes an account on the social media channel, while *smio:Channel* represents a social media plattform e.g. Twitter. *smio:Post* is a text message published by users. *smio:POS* describes a superclass for different types of words (Part-of-Speech) and contains the subclasses. *smio:ProperNoun* (proprietary name), *smio:CommonNoun* (No-men), *smio:Adjective* and Triplet. A triplet is the upper class of *smio: Subject*, *smio:Verb* and *smio:Object*. It is assigned to subclasses of the *smio:POS* class. The *rdf: Bag* class is a container for unordered lists and is used for keywords of the tweet. *smio:CustomerNeed* describes a superclass for different types of customer needs e.g. a desired feature for a product or a negative experience with a product. This superclass contains the subclasses *smio:Feature*, with words related to the product like specific parts of a product ”amoled display” and negative *smio:Experience*, with words like ”annoyed”, ”irreparable”, ”boring”, ”unsightly”, ”broken”). The *smio:Category* superclass describes the kind of Category a Customer Need is related to. This superclass contains the subclasses *smio:Produkt*, *smio:Service* and *smio:Process*. Class *smio:Characteristic* specifies the part of Customer need. Wheter it is a customer need described in the form of a *smio:Problem* (Word like ”unremedied”, ”trouble”, ”issue”, ”hitch”, ”struggle”) or it is a solution related to a customer need *smio:Characteristic* (Characteriesd by words like ”solved”, ”fixed”, ”remedied”). An *smio:NamedEntity* serves as a class for identified categories of proper names by Named Entity Recognition (NER).

2) Properties: The *smio:submitted* property is associated with a *smio:User* and a *smio:Post*. The property *smio:containsPos* is assigned to a *smio:Post* and a subclass of the *smio:POS* class. The property *smio:hasKeywords* is assigned to a *smio:Post* and a *rdf:Bag*. The *smio:containsCustomerNeed* is associated with a *smio:Post* and *smio:CustomerNeed*. *smio:relateToCategory* is assigned to a *smio:CustomerNeed* and *smio:Category*. The *smio:relateToEntity* property is associated with a *smio:CustomerNeed* and a *smio:NamedEntity*, while property *smio:hasCharacteristic* is associated with a *smio:CustomerNeed* and *smio:Characteristic*.

3) Literals: The literal *smio:userID* is assigned to the *smio:User* class and describes the unique identifier of the user. *smio:userName* is assigned to class *smio:User* and describes the user name of the social media account e.g. Twitter. The literal *smio:reputation* is assigned to class *smio:User* and reflects a user’s reputation in the form of a number between 0 and one. As shown in a previous research project, it is

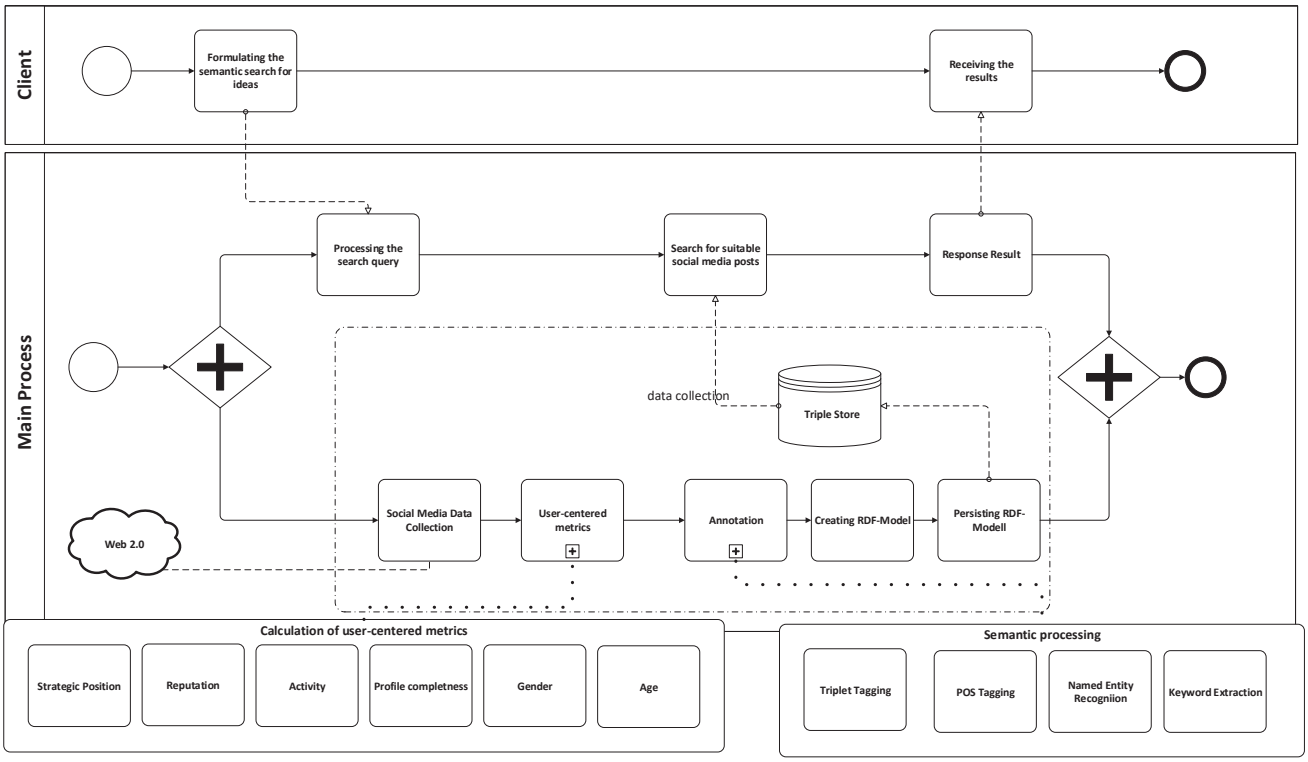

Fig. 1 Overview of the Concept

possible to measure the reputation of a user in a social media channel. For this purpose, metrics based on the amount of input nodes and messages related to the user as well as social media channel-specific reputation indicators were developed [16]. The literal *smio:activity* is assigned to class *smio:User* and indicates how active a user is within the network. As mentioned in [16], the activity of a user can be measured through the ratio of the age of the account in days to the number of composed messages. If a user composes many messages and his account is young, the value converges towards 1. The literal *smio:position* is assigned to class *smio:User* and indicates the strategic position of a user. According to Eccleston and Griseri [17], this metric shows how well a user is embedded in a social network. For example, a SPAM bot has a lower strategic position than an influencer. The literal *smio:completendessInfo* is assigned to the class *smio:User* and shows, according to [4], whether the user is authenticated in the network or not. This is measured, for example, by the completeness of the profile information. The literals *smio:gender* and *smio:age* are assigned to the class *smio: User*.The information about gender is determined via the nameapi-API [18] based on the users name. Our approach based on the assumption of Rao et al. [19] and classified user into classes up to 30 years and over 30 years. The chance to determine the age and gender of a user depends heavily on the quality of the user profile and the written text and is not always possible.

V. Tagging

A. Triple Taggers

Three possible variants of triplet taggers were compared for the extraction of the subject verb object triplets (SVO).

1) Stanford Parser: In this variant, an algorithm for identifying the triplets has been implemented, which uses the so-called Typed Dependencies of the Stanford CoreNLP Java Suite. Typed Dependencies is an easy-to-understand description of the grammatical relationships of words in a sentence. These are generated by the Stanford parser, which is a probabilistic parser that performs the syntactic analysis of unstructured text using control probabilities [20].

2) IBM Watson Natural Language Understanding API: The IBM Watson Natural Language Understanding API uses a programming interface from IBM's Watson cognitive software, which processes the text using proprietary algorithms. Watson offers numerous possibilities for processing natural language and text analysis.

3) Stanford Open Information Extraction: Triplets are extracted in Stanford Open Information Extraction (Stanford OpenIE) using an alternative method that, unlike the Stanford parser, uses significantly fewer predefined grammatical patterns and can therefore be used in many application scenarios [21]. This also applies to tweets, for example, which often contain short sentences with erroneous sentence structure, which often pushes rigid predefined patterns to their limits.

4) Evaluation of the Triplet Taggers: To decide which of the listed taggers is most suitable for Twitter data detection, a gold standard for the taggers was determined. For this purpose,

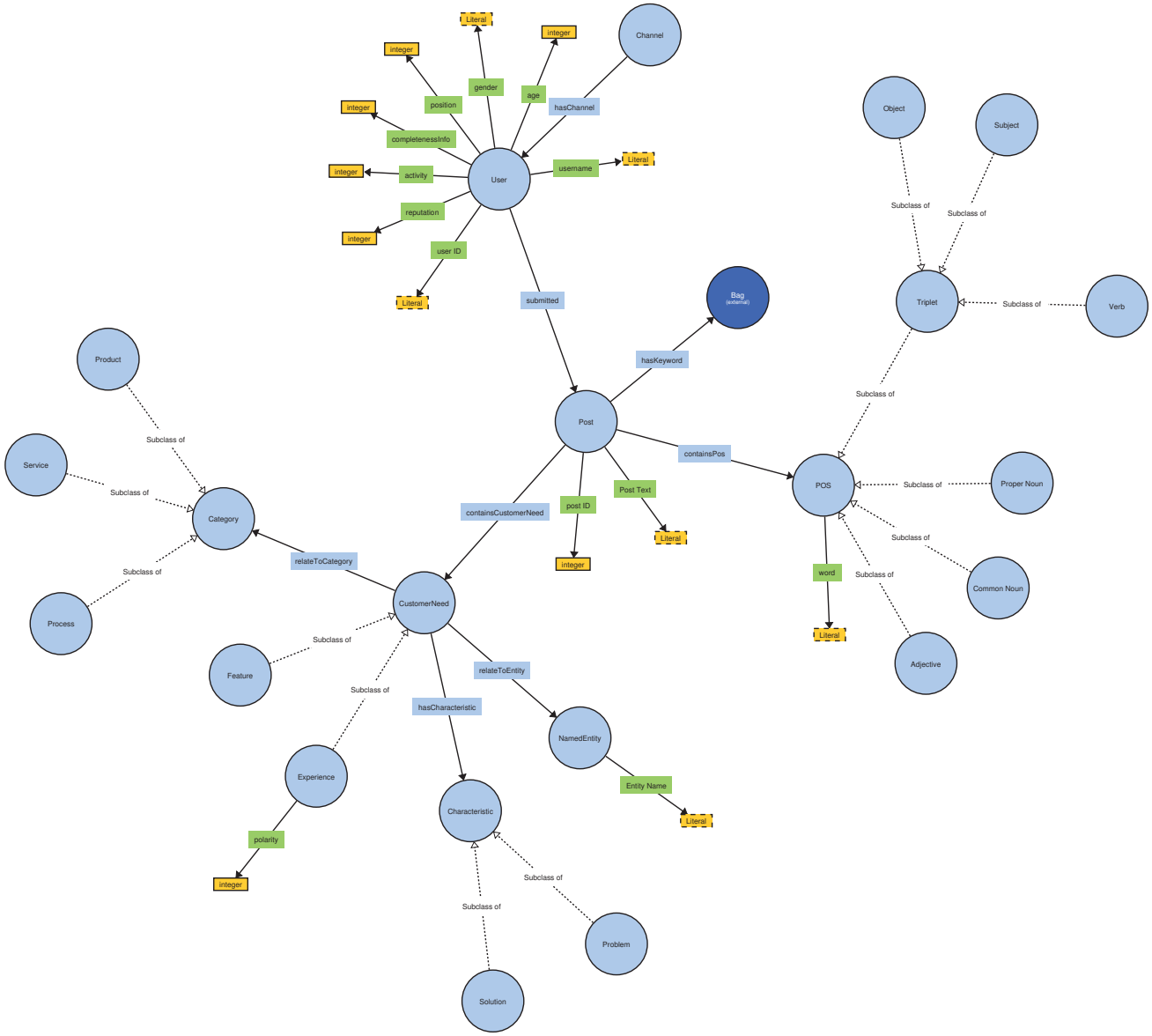

Fig. 2 Social Media Idea Ontology (SMIO)

333 random tweets were extracted and manually annotated their subject, verb and object. The manually annotated tweets were then also examined by the three taggers described above.

The *Stanford Open Information Extraction* extracts as many SVO triplets as possible. For example, up to eight triplets were detected by OpenIE for some manually detected triplets. Due to this high number of SVO triplets, all information is recognized in most cases, but the number of incorrectly detected triplets is also increasing rapidly. In addition, auxiliary verbs such as “is”, “be” or “are” are often used. This leads to false triplets being detected, especially with compound expressions such as Machine Learning.

The *WatsonNLU* always tries to extract the subject, verb and object that has been possible for a long time. Thus, adjectives and adverbs are extracted with the subject or object. This enables a later full text search without loss of information. Similar to the OpenIE Tagger, Watson NLU also has weaknesses with auxiliary verbs in compound expressions.

The *Stanford Parser*, on the other hand, often extracts only one word for the subject or object, in contrast to Watson. Especially for compound expressions, this leads to incomplete parts being taken over, which also changes the meaning.

The taggers are evaluated using recall, precision and F1 values. The recall value specifies which portion of the manually tagged values was recognized. The precision value specifies which portion of the values extracted by a tagger is correct. The F1 value is used to give a summary overview and to adjust precision and recall values. Table 1 shows the values for the three taggers.

Based on these results, we choose WatsonNLU.

B. Part-of-Speech-Tagging

Although such systems have been around for some time, tweets often achieve only moderate results due to numerous

TABLE I
RESULT OF THE TRIPLET TAGGER EVALUATION

Tagger	Recall	Precission	F1
OpenIE	54,4%	62,6%	58,2%
WatsonNLU	56,2%	72,2%	63,2%
Stanford Parser	19,5%	46,4%	27,5%

linguistic errors and idiosyncratic terms. The individual types of words, such as adjectives or nouns, are only insufficiently extracted. For this reason, the results of Derczynski et al. [22] were used in this work. The authors have developed an optimized model for the classification of English-language tweets and have thus achieved the highest accuracy of tweets classified with the Penn Treebank Tagset [23] to date. The model is used in conjunction with the Stanford POS-Tagger [24] and has been integrated into the StanfordTagger class application.

C. Named Entity Recognition

Named Entity Recognition (NER, also known as Named Entity Extraction) aims to identify the entities contained in a text. Historically, the term entity has been used since the 1970s, especially in the field of data modelling - the entity relationship model. An entity is a unique object of the real world (e. g. the person John Doe) about which information is to be stored or processed [25]. A common NER technique is the usage of Hidden Markov Models [26] . To use our defined dictonaries to annotate the text by *Feature*, *Experience* or *Characteristic*, we use the cloud-service discussed above. The difference to the already know *Named Entities* is, the training of the cloud-service. This we did by annotating 100 significant Social Media Posts for each Kategorie. To identify and annotate the related category of a post, we used choosen cloud-service as well.

D. Keyword Exctraction

In contrast to Named Entity Recognition, Keyword Extraction (KE) does not extract the proper names, but extracts all relevant information from a text. These may also be proper names as well as concepts or quotations. These key sentences or words are the terms that best characterize a text by its content. The aim of a keyword extraction is to condense a text and extract the relevant information from it [27]. TF-IDF is a simple but widely used numerical statistic to extract keywords [28].

There are several companies that offer the above-mentioned NLP procedures as a cloud-service via the internet. Basically, a distinction must be made between classic applications and cloud-services. Within the scope of this research, the focus lies on the cloud-services, which are provided by several companies. In this research work the three cloud-service (APIs) WatsonNLU, DandelionAPI and MeaningCloud were chosen to be evaluated after a pre-research of eight cloud services. The APIs have been selected by the difference in the providing companies. While a small Italian start-up is responsible for the development of DandelionAPI, Watson NLU is operated by the multi-billion-dollar IBM group, while MeaningCloud Service is operated by a medium-sized company.

To decide which of the cloud-service is most suitable for our case a manually labelled gold standard with 500 tweets was created. In order to compare cloud-services, the metrics Precision and Recall as well as the harmonic mean (F1) and Accuracy were calculated. True positive (TP) was evaluated if the gold standard contained the entity and the API also classified it as an entity. A false negative (FN) was evaluated if the gold standard contained the entity, but the API did not classify it as an entity. The false positive (FP) was considered false if the gold standard did not contain the entity, but the API classified it as an entity. If the gold standard contained the entity and the API did not classify the word as an entity, this was considered true negative (TN). An overview of the result is given in table 2.

TABLE II
RESULT OF CLOUD-SERVICE EVALUATION

Cloud-Service	Recall	Precission	F1	Accuracy
MeaningCloud	28,82%	38,2%	39,12%	63,29%
Dandelion	53,09%	53,40%	57,05%	72,77%
WatsonNLU	77,54%	52,61%	57,00%	79,71%

Based on these results, we choose Dandelion.

VI. PRE-STUDY RESULTS AND OUTLOOK

In this paper we showed the evaluation concepts of the Taggers and the results in Section V. In addition, Fig. 3 shows the current status of our user interface conceptually, followed by a brief explanation of the functions. The contextual restrictions can be set in the User and Social Media Post area and marked with number one. In the case of the user area, this is done by selecting and entering thresholds, and in the case of social media posts by selecting included or excluded properties. An example of this would be the specific requirement that the searched *subject* also has to be a *Proper Noun*. Within the *Output Filter*, marked with number two, the user can specify which fields are to be displayed in the later result, that is, which columns are to be displayed in the table. By selecting the *search button* marked with number three, the SARQL query is generated and passed to the Fuseki server. The server then returns the result to the front-end where it is displayed in the result pane (marked with number four) with the fields selected in the Output Filter pane.

In order to evaluate the user-centered metrics and to determine the exact search pattern, we have recently launched an online survey. After evaluation and incorporation of the results, a further evaluation of the overall concept and thus the ontology will follow. A domain-specific test with product experts is being planned. In an initial experimental test, many features of a given consumer product have already been identified. The next step is to improve the automatic annotation. For this purpose, the annotation based on dictionaries is intended to convert a machine learning based annotation.

VII. CONCLUSION

The advantage of our concept lies in the focus on public social media channels and the strong inclusion of user-centered

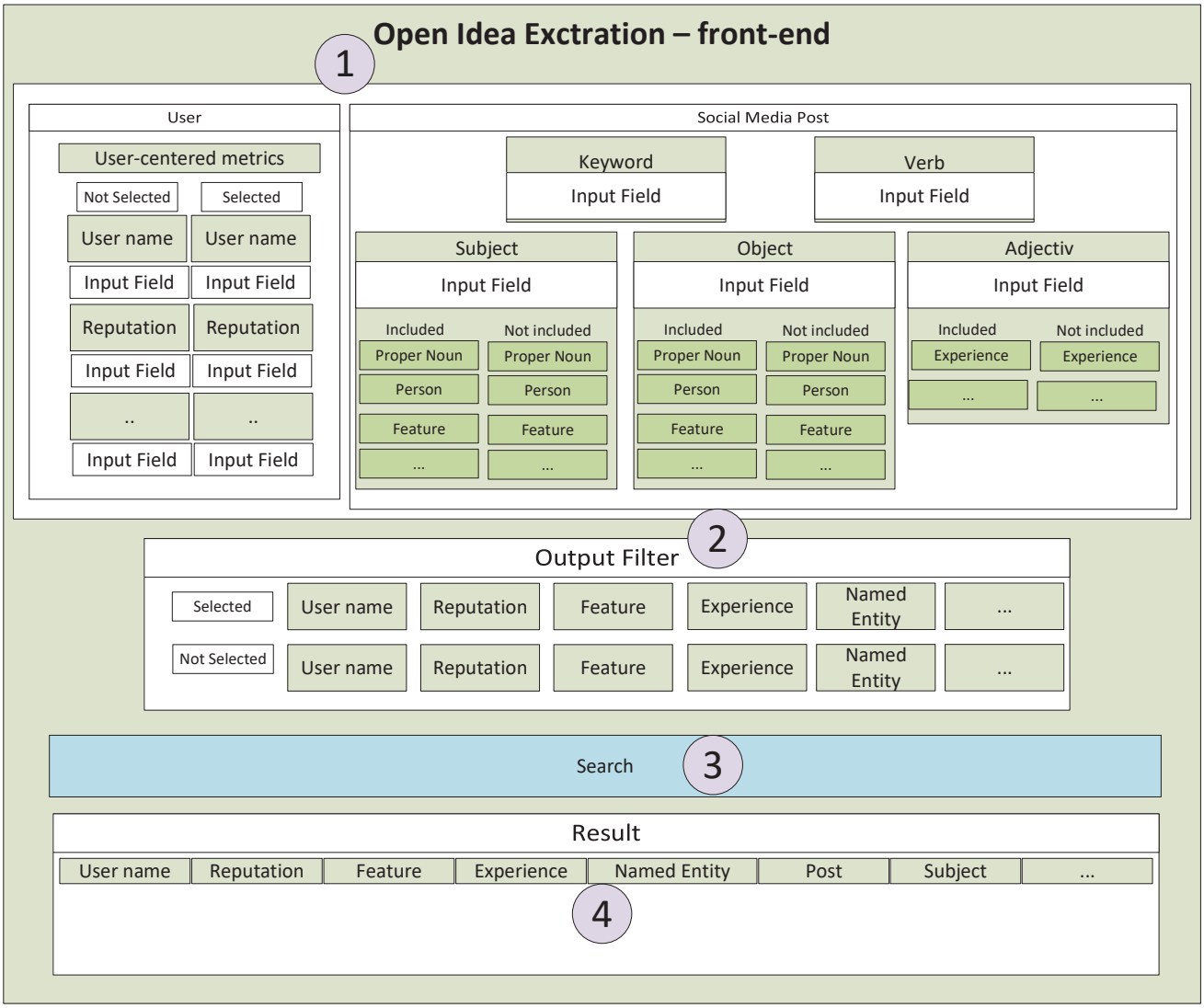

Fig. 3 Front end to search within social media data

metrics. Thus, these metrics and the semantic processing make it possible to quickly and purposefully search for potential ideas for products in a large amount of user-generated data.

REFERENCES

[1] H. Smith, "A CSC white paper, european office of technology and innovation. what innovation is. how companies develop operating systems for innovation." (Online). Available: http://www.innovationmanagement.se/wp-content/uploads/pdf/innovation_update_2005.pdf, last accessed on 11.11.2017.

[2] J. Xu, R. Houssin, E. Caillaud, and M. Gardoni, "Macro process of knowledge management for continuous innovation," in *Journal of Knowledge Management*, vol. 14, pp. 573–591. (Online). Available: http://www.emeraldinsight.com/doi/abs/10.1108/13673271011059536, last accessed on 03.04.2017.

[3] A.-L. Mention, "Co-operation and co-opetition as open innovation practices in the service sector: Which influence on innovation novelty?" in *Technovation*, ser. Open Innovation - ISPIM Selected Papers, vol. 31, pp. 44–53. (Online). Available: http://www.sciencedirect.com/science/article/pii/S0166497210000908, last accessed on 03.08.2017.

[4] R. Alt and O. Reinhold, *Social Customer Relationship Management*. Springer Berlin Heidelberg. (Online). Available: http://link.springer.com/10.1007/978-3-662-52790-0, last accessed on 22.01.2017.

[5] H. I. Ansoff, "Managing strategic surprise by response to weak signals," vol. 18, no. 2, pp. 21–33. (Online). Available: https://doi.org/10.2307/41164635 last accessed on 07.010.2017.

[6] R. Eckhoff, J. Frank, M. Markus, M. Lassnig, and S. Schoen, "Detecting innovation signals with technology-enhanced social media analysis - experiences with a hybrid approach in three branches," vol. 17, no. 1, pp. 120–130. (Online). Available: http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-15-065-09, last accessed on 07.06.2017.

[7] M. Markus, R. A. Eckhoff, and M. Lassnig, "Innovation signals in online-communitys ein komplementaerer analytischer ansatz," vol. 50, no. 5, pp. 13–21. (Online). Available: http://link.springer.com/10.1007/BF03340849, last accessed on 17.11.2017.

[8] Apache tomcat (Online). Available: http://tomcat.apache.org/, last accessed on 17.11.2017.

[9] Apache marmotta. (Online). Available: http://marmotta.apache.org/, last accessed on 17.11.2017.

[10] Apache marmotta - KiWi triple store. (Online). Available: http://marmotta.apache.org/kiwi/, last accessed on 17.11.2017.

[11] SPARQL query language for RDF. (Online). Available: https://www.w3.org/TR/rdf-sparql-query/, last accessed on 17.11.2017.

[12] D. Thorleuchter, D. V. den Poel, and A. Prinzie, "Mining ideas from textual information," vol. 37, no. 10, pp. 7182–7188. (Online). Available:

http://linkinghub.elsevier.com/retrieve/pii/S0957417410002848, last accessed on 29.10.2017.

[13] D. Thorleuchter and D. Van den Poel, "Web mining based extraction of problem solution ideas," vol. 40, no. 10, pp. 3961–3969. (Online). Available: http://linkinghub.elsevier.com/retrieve/pii/S095741741300016X, last accessed on 28.10.2017.

[14] A. Westerski, C. A. Iglesias, and F. T. Rico, "A model for integration and interlinking of idea management systems," in *Metadata and Semantic Research*, ser. Communications in Computer and Information Science. Springer, Berlin, Heidelberg, pp. 183–194. (Online). Available: https://link.springer.com/chapter/10.1007/978-3-642-16552-8_18, last accessed on 17.10.2017.

[15] Idea storm. (Online). Available: http://www.ideastorm.com/, last accessed on 20.10.2017.

[16] M. Haeusl, J. Forster, and D. Kailer, "An approach to identify SPAM tweets based on metadata." IEEE, pp. 48–51. (Online). Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7397420, last accessed on 17.03.2017.

[17] Eccleston and Griseri, "How does web 2.0 stretch traditional influencing patterns? international journal of market research," no. 50, pp. 591–661.

[18] NameAPI - intelligence in names. (Online). Available: https://www.nameapi.org/, last accessed on 22.10.2017.

[19] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter." ACM Press, p. 37. (Online). Available: http://portal.acm.org/citation.cfm?doid=1871985.1871993, last accessed on 20.10.2017.

[20] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Association for Computational Linguistics, pp. 423–430. (Online). Available: https://doi.org/10.3115/1075096.1075150, last accessed on 08.11.2017.

[21] G. Angeli, M. Jose Johnson Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," vol. 1, pp. 344–354.

[22] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," in *International Conference Recent Advances in Natural Language Processing, RANLP*.

[23] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," vol. 19, no. 2, pp. 313–330. (Online). Available: http://dl.acm.org/citation.cfm?id=972470.972475, last accessed on 12.10.2017.

[24] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Association for Computational Linguistics, pp. 173–180. (Online). Available: https://doi.org/10.3115/1073445.1073478, last accessed on 01.11.2017.

[25] P. P.-S. Chen, "The entity-relationship model toward a unified view of data," vol. 1, no. 1, pp. 9–36. (Online). Available: http://doi.acm.org/10.1145/320434.320440, last accessed on 05.11.2017.

[26] B. T. Todorovic, S. R. Rancic, I. M. Markovic, E. H. Mulalic, and V. M. Ilic, "Named entity recognition and classification using context hidden markov model," in *2008 9th Symposium on Neural Network Applications in Electrical Engineering*, pp. 43–46.

[27] P. D. Turney, "Learning algorithms for keyphrase extraction," vol. 2, no. 4, pp. 303–336. (Online). Available: http://link.springer.com/article/10.1023/A:1009976227802, last accessed on 01.11.2017.

[28] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.