Widen the Peepholes! Entity-based Auto-Suggestion as a rich and yet immediate Starting Point for Exploratory Search

Johannes Osterhoff, Jörg Waitelonis, and Harald Sack
Hasso Plattner Institute for IT Systems Engineering
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

{johannes.osterhoff, joerg.waitelonis,
harald.sack}@hpi.uni-potsdam.de

Abstract: Today's search engines provide instant keyword-based auto-suggestion and completion of the user's search queries. This paper presents a novel auto-suggestion interface for the Semantic Multimedia Explorer (SEMEX), a semantic search engine that supports entity-based exploratory video retrieval. In difference to traditional text-based retrieval, auto-suggestion and auto-completion of the user's query string is not based on plain text but on semantic entities grouped by meaningful categories. Suggested entities are ranked by plain edit-distance as well as by popularity while their representations are enabled for brushing and linking. The categories contribute to a quick comprehensibility of the suggested entities compared to other systems that constrain an actually significant and practical feature into static and narrow vertical listings. Thus, our approach is leading to better decision making from the very start of an exploratory search.

1 Introduction

Today's web search engines facilitate the formulation of search queries by providing instant keyword-based auto-suggestion and auto-completion of query strings. Popular search engines such as Google[1] enable suggestions by means of simple distance measure based on edit distance and popularity in the query log. Usually displayed as a handful of suggestions in a flat vertical listing, this drop-down box has never been more than a peephole to the vast amount of search possibilities. Since most recent time, the extraction of semantic entities from huge data collections as well as to find meaningful representations on graphical interface level is understood to be one of the most important challenges in search technology [BYBM11].

This so-called *semantic search* typically is based on semantic entities, enabling the user to make more precise decisions during the search process. Furthermore, faceted refinement and content-based suggestions are used to drill down and broaden search queries leading to exploratory search [WKW+10]. While semantic entities can also be used to expand

[1] http://www.google.com/

the user query terms [TS10] and query a traditional search engine, this work focusses on systems providing genuine entity based semantic search.

Semantic entity-based search is capable to overrule keyword-based search in preciseness and recall ratio because of improvements, as e. g. in relevance based document ranking, which now can incorporate semantic relations between the indexed documents [BMdR10]. Furthermore, semantic search is robust against natural language phenomena, such as polysemy and synonymity, which causes ambiguities in search results. For the disambiguation of a semantic query term the user can be supported by auto-suggestion methods providing a reasonable selection of meaningful semantic entities fitting the given text input. Following the same principle, semantic entity-based auto-suggestion can also be applied for semantic document annotation. Therefore, it is an important prerequisite to enable semantic aware text processing, as e. g., in semantic Wiki-systems, semantic tagging services, or semantic content management systems [LS09].

Especially in the context of query string refinement and completion, the simple visual representation of traditional auto-suggestion has to be reconsidered to be more than a peephole for expected results, but a useful tool supporting the user's exploratory decision making process from the very start.

Since each suggested semantic entity might belong to several ontological classes, an objective way has to be found to make use of these structures without suggesting a misleading emphasis to the user. In addition, in comparison to keyword-based suggestions, it is not apparent why certain entities are displayed, since the reasons go beyond straightforward visual comparability. For example, a semantic entity might be suggested because it is a synonym of the query string or it might match several different categories. Semantic auto-suggestion also is expected to reveal meaningful relations of the suggestions with each other, making it possible for the user to compare the displayed entities and relate them to each other, allowing a precise and conscious selection.

Taking these aspects into consideration, this paper illustrates an efficient semantic entity-based auto-suggestion interface for the semantic video search engine SEMEX[2]. The paper is structured as follows: In Sect. 2 related work is described and a relevant interface paradigm is introduced. Sect. 3 deals with the realization of the auto-suggestion including design aspects, adaptations for touch environments, and indexing algorithms. Finally, Sect. 4 concludes the paper with a short discussion of possible ways to for evaluation and already achieved results.

2 Related Work

Auto-suggestion as well as auto-completion is a mechanism in which, as users enter a search term into a search box, related queries are shown below [BW06]. This attempt to help users finish entering their queries is understood to be of high usability in general

[2]SEMEX has been developed during the Mediaglobe project. Mediaglobe is a SME project of the THESEUS research program, supported by the German Federal Ministry of Economics and Technology on the basis of a decision by the German Bundestag.

[WM07, Hea08]. Usually suggestions are provided in drop-down boxes and list keywords that have been provided by other users in previous searches. In the context of semantic entity-based search auto-suggestion can and has been used to display more than just keyword text strings, leading to more complex layouts of the auto-suggestion interfaces: Freebase[3], a database of structured data harvested from various sources, makes use of an entity-based auto-suggestion mechanism implemented as scrollable drop-down. Titles are presented as primary list entries while their appropriate semantic categories are shown on the right. The entities are neither grouped nor aggregated, but for each entity a preview in a pop-up window is provided, which contains a thumbnail and an introductory text snippet. This pop-up window offers a detailed content-based preview, which is offered by the proposed auto-suggestion only after an explicit selection. The introductory text of Freebase guides the user to select a suitable entity, but it does not facilitate the user to easily compare entities among each other by visual feedback. Huynh et al. introduced the Parallax navigation interface for Freebase data [HK09], which allows navigation of this structured data mainly along facets. The auto-suggestion mechanism of this interface is subdivided into topics mentioning the search term in their text context and individual topics resembling it. In the latter, semantic entities and labels are listed. The interface of the cultural search engine MultimediaN [SA+08] also makes use of a vertical drop-down for auto-suggestion. In this example, each of the semantic entities are attributed with only one class and categorized in persons, locations, artifacts, concepts and others. In each category three results are shown but this list can be expanded to list all available suggestions. The proposed auto-suggestion provides more space to its suggestions compared to the Parallax navigation interface and MultimediaN. The layout is arranged in a more spacious and a horizontally way. After pressing a button to view more suggestions on both interfaces, the vertical scrolling in a rather narrow visible area impairs the clarity of the listed suggestions. The Finnish cultural search engine CultureSampo provides several interfaces for faceted semantic recommendations, organizing places, people, and relations from a collaboratively generated ontology [HM+09]. Its Quick Search makes use of the entire screen for its disambiguation and presents semantic entities of distinct categories together in one vertical listing. For each entity, a selection of appropriate semantic categories is given and entities may be distinguished by means of category icons. In case a general search query is entered, the listing of CultureSampo tends to become very long and is apparent that a vertical division of the layout would provide a better overview to the presentation of suggestions since unnecessary scrolling would be avoided. Also concerned with the exploratory aspect of auto-suggestion is the recent SparQS interface [KST12] by Kato et al. This interface was developed to facilitate its users, both to specialize their queries, as well as to contribute to their "parallel movement", which allows to switch to another topic of interest with similar aspects. In this example, the combination of instant refinement and exploration is provided by entities as alternatives to the currently suggested entities aligned in a grouped tab-like vertical listing. Such a layout clearly structures the suggestions, but it also deems specialization more important than exploration. The layout of the SEMEX auto-suggestion prevents such an emphasis and displays all suggested entities at par. In addition a vertical layout might be difficult to read, especially when it comes to internationalization with non-latin typefaces.

[3] http://www.freebase.com/

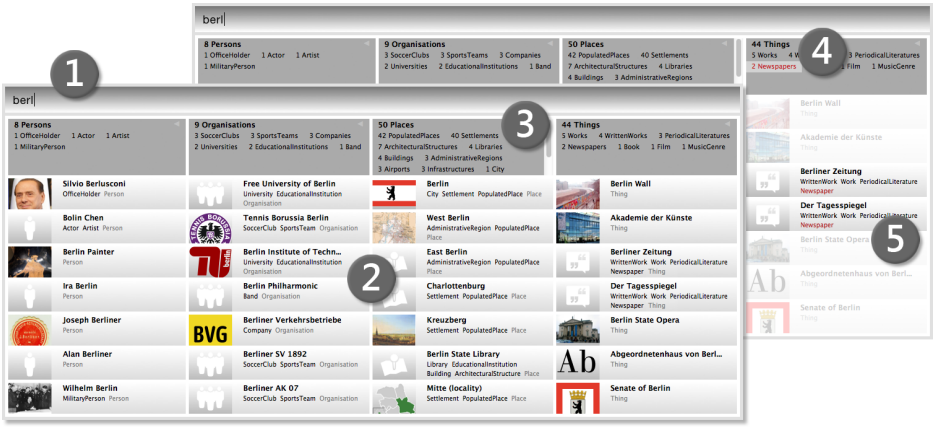

Figure 1: SEMEX auto-suggestion with semantic categories in its header and columns of suggested entities. For better readability the column presenting events has been omitted.

The principle of "Brushing and Linking", which is used for the proposed auto-suggestion, and also the presentation of search results of SEMEX [OW+11] originates from early experiments in computer graphics and has become a common method in information visualization. Brushing and linking describes a connection between two or more views of the same data in a way that a change to the representation in one of the views also affects the representation in the other ones. The principle was first introduced by Becker and Cleveland [BC87] to brush and link values of scatterplot matrices in the late 1980s. The proposed auto-suggestion facilitates users to *brush* semantic classes listed in the captions of each category and *links* them to the actual suggestions.

3 The SEMEX Auto-Suggestion

Considering Fig. 1, the layout of the proposed auto-suggestion is divided into a search box (1) and a disambiguation matrix (2).[4] While typing a query string into the search mask, the disambiguation matrix shows up. This matrix spreads over the whole width of the layout and is vertically subdivided into the five categories Persons, Organizations, Places, Events, and Things.

During text input these segments update immediately according to user input and show relevant semantic entities. In case no suggestions occur in a specific category, this column is not displayed. Each of these entity suggestions comprises a title and a subtitle in which the entity's semantic categories are displayed. If available, a thumbnail image originating from Wikipedia's Commons is prepended. On top of each column, reside aggregated semantic categories of the entities below (3). These captions are ordered by occurrence in

[4] A live demo of the auto-suggestion can be found at http://www.yovisto.com/labs/autosuggestion/

the corresponding section and thus add to the comprehensibility of the auto-suggestion. In addition, the captions are enabled by brushing and linking: when the user selects a caption (4), all entities that bear the same semantic category are highlighted (5). This highlighting of suggestions offers a quick comparability of entities upon user interaction by brushing. In addition, brushing also offers a simple way to undo a selection – quicker than for example explicit filtering with an refreshing of listed suggestions. In case a category in the subtitle of an entity is selected, again other appropriate entities are highlighted. When selecting an entity suggestion by its title, the auto-suggestion is closed and a new search is performed.

The entity suggestions are based on the DBpedia[5] datasets. Every entity is indexed via unique URI, a main label, the DBpedia ontology classes the entity belongs to, and a list of related labels generated from DBpedia redirects. The related labels include alternate spellings, synonym spellings, misspellings, and other descriptive labels. For every manually selected category (Persons, Organizations, Places, Events, and Things) a separate Lucene[6] index is generated to query each category individually for performance reasons. We have selected these categories under the assumption that users are mainly interested in items of these types.

The suggestions for a given query string have to be ranked appropriately to support the user surveying all suggestions at a glance. Matches are presented in the following order: exact matches, matching words, labels with matching prefix, and labels with matching sub-string. Furthermore, entity popularity should also be included to ensure the suggestion ranking meets the most common user expectations. The TF/IDF scoring applied in traditional information retrieval [BYRN99] is not appropriate to rank the semantic entities, because entities are not structured like text documents. In this application, term frequency (TF) is not nessessarily an indicator of high relevance. Entities can have a totally different number of alternate labels containing different spellings and writings (e. g. `dbp:Berlin_tram` entails more synonyms than `dbp:Berlin`) which would have the effect to boost entities with a higher number of alternate spellings.

Instead of TF/IDF, the proposed ranking is based on a string distance measurement between the label h, which contains the search hit and the main label l of the entity. The $score$ is determined as:

$$score(l,h) = \begin{cases} 1.0, & exact\ match \\ r, & word\ match \\ r * JaroWinkler(l,h), & prefix\ match \\ r^2 * JaroWinkler(l,h), & else, \end{cases}$$

where $0 < r < 1$. We have chosen $r = 0.9$ empirically. Incorporating the general popularity of entities in the final rank is achieved by subsequently ordering the top $n = 50$ scoring semantic entities according to the number of incoming internal Wikipedia links of the entity's corresponding Wikipedia article. Thereby, the principle of link popularity is applied, which we consider as indicator of commonly accepted popularity rating.

Interaction on touch-enabled surfaces is more direct compared to traditional desktop environments [FW+07]. Due to the nature of this emerging form of human computer inter-

[5] http://dbpedia.org/

[6] http://lucene.apache.org/

action, the representation of a mouse pointer increasingly is abandoned for a more direct manipulation of the objects on the surface of the screen by the hand itself. And yet, the subtlety of brushing objects without explicitly selecting them, which is required for the proposed auto-suggestion, might be lost on tablet computers. To make use of the proposed interaction paradigm, classical brushing and linking with a mouse pointer has to be replaced with a meaningful manipulation vocabulary for touch. To adapt the proposed auto-suggestion of SEMEX for touch environments, the gestures for brushing and the definite selection have been approximated. Users might select a caption by tap and see the highlighting of the fitting entity suggestions. In case users choose to select a facet permanently, they will have to tap the entities' title.

4 Conclusion and Future Work

The capabilities of the proposed SEMEX auto-suggestion cannot be achieved by traditional text-based auto-suggestions. The aggregation of DBpedia categories as captions for suggested entities and their vertical alignment offers better clarity and comprehensibility. The application of brushing and linking to these captions, as well as to entities' subtitles, contributes to an exploration of the search space from the very start.

There are still some open issues and improvements to be handled in future work, e. g. the suggestions only display the main label of the entity, which sometimes irritates the user – she cannot verify why a suggestion has been selected as result. This happens for example, if a hit occurs in a synonym. It is even more unclear, when in this synonym term only a substring match occurs. Furthermore, it should be considered not only to use selected properties of an entity, but incorporate all textual properties.

Due to the focus of user interface design, the entity ranking methods are not evaluated in this paper and remain future work. To evaluate the SEMEX auto-suggestion interface we plan to conduct a user study, whose participants will be asked to fulfill a set of tasks within a specific time-frame. Participants would then either use the SEMEX auto-suggestion or a flat vertical listing for comparison. To evaluate the users' acceptance of the exploratory brushing and linking feature, we plan to log its overall usage during the study without letting participants know about it. A change of usage over time, then would show the acceptance or rejection of such a feature by its users.

Future work also might investigate, if the brushing of subcategories will affect all other categories. For example, selecting "soccer club" in the Organization category might result in highlighting "soccer player" in Persons, because there is a relationship between these two concepts. Besides it might be useful, if a user selects a subcategory that the search is repeated but restricted to that subcategory. Thus, users might refine their search and see more results matching their query. In addition the interface might be enriched with a live preview of the result set, which immediately updates upon brushing in the auto-suggestion.

References

[BC87] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29:127–142, May 1987.

[BMdR10] K. Balog, E. Meij, and M. de Rijke. Entity Search: Building Bridges between Two Worlds. In *SemSearch2010: Semantic Search 2010 Workshop at WWW 2010*, Raleigh, NC, 2010. ACM.

[BW06] H. Bast and I. Weber. Type Less, Find More: Fast Autocompletion Search with a Succinct Index. In *Proc. of the 29th annual Int. ACM SIGIR Conf. on Research and development in information retrieval*, pages 364–371, New York, NY, USA, 2006. ACM.

[BYBM11] R. Baeza-Yates, A. Broder, and Y. Maarek. The New Frontier of Web Search Technology: Seven Challenges. In S. Ceri and M. Brambilla, editors, *Search Computing*, volume 6585 of *LNCS*, pages 3–9. Springer Berlin / Heidelberg, 2011.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[FW+07] C. Forlines, D. Wigdor, et al. Direct-touch vs. Mouse Input for Tabletop Displays. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 647–656, New York, NY, USA, 2007. ACM.

[Hea08] M. A. Hearst. UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In *HCIR08 2nd Workshop on Human-Computer Interaction and Information Retrieval*. Microsoft, October 2008.

[HK09] D. Huynh and D. Karger. Parallax and Companion: Set-based Browsing for the Data Web. 2009.

[HM+09] E. Hyvönen, E. Mäkelä, et al. CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0. In *Proc. of 6th European Semantic Web Conference (ESWC 2009)*, pages 851–856. Springer Berlin / Heidelberg, 2009.

[KST12] M. A. Kato, T. Sakai, and K. Tanaka. Structured Query Suggestion for Specialization and Parallel Movement: Effect on Search Behaviors. In *Proc of the 21st Int. Conf. on World Wide Web (WWW 2012)*, New York, NY, USA, 2012. ACM.

[LS09] R. Landefeld and H. Sack. Collaborative Web-Publishing with a Semantic Wiki. In S. Schaffert, T. Pellegrini, et al., editors, *Networked Knowledge – Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems*, pages 129–140. Springer Berlin / Heidelberg, 2009.

[OW+11] J. Osterhoff, J. Waitelonis, et al. Sneak Preview? Instantly Know What To Expect In Faceted Browsing. In *Proc. of Workshop Interaktion und Visualisierung im Daten-Web*, 2011.

[SA+08] G. Schreiber, A. Amin, et al. Semantic Annotation and Search of Cultural-heritage Collections: The MultimediaN E-Culture Demonstrator. *Web Semant.*, 6(4):243–249, November 2008.

[TS10] S. L. Tomassen and D. Strasunskas. Constructing Feature Vectors for Search: Investigating Intrinsic Quality Impact on Search Performance. *Int. J. Web Grid Serv.*, 6(3):289–312, September 2010.

[WKW+10] J. Waitelonis, M. Knuth, L. Wolf, et al. The Path is the Destination – Enabling a New Search Paradigm with Linked Data. In *Proc. of the Workshop on Linked Data in the Future Internet at the Future Internet Assembly, Dec. 16-17, 2010, Ghent, Belgium, CEUR Workshop Proc.*, volume 700, 2010.

[WM07] R. W. White and G. Marchionini. Examining the Effectiveness of Real-time Query Expansion. *Inf. Process. Manage.*, 43(3):685–704, May 2007.