Visual Analytics and Similarity Search: Concepts and Challenges for Effective Retrieval Considering Users, Tasks, and Data

Daniel Seebacher[1], Johannes Häußler[1], Manuel Stein[1],
Halldor Janetzko[2], Tobias Schreck[3], and Daniel A. Keim[1]

[1] University of Konstanz, lastname@dbvis.inf.uni-konstanz.de
[2] University of Zurich, halldor.janetzko@geo.uzh.ch
[3] TU Graz, tobias.schreck@cgv.tugraz.at

Abstract. A major challenge of the contemporary information age is the overwhelming and increasing data amount, especially when looking for specific information. Searching for relevant information is no longer manually possible, but has to rely on automatic methods, specifically, similarity search. From a formal perspective, similarity search can be seen as the problem of finding entities, which are considered to be similar to a query with respect to certain describing features. The question which features or which weighted combination of features to use for a given query creates a need for semi-automatic methods to address the needs of diverse users. Furthermore, the quality of the results of a similarity search is more than effectiveness, measured by precision and recall. The user ideally needs to trust the results and understand how they were computed. We propose to apply Visual Analytics methodologies, for synergistic cooperation of user and algorithms, to integrate three key dimensions of similarity search: users, tasks, and data for effective search. However, there exists a gap in knowledge how user, task as well as the available data influence each other and the similarity search. In this concept paper, we envision how Visual Analytics can be used to tackle current challenges of similarity search.

Keywords: Similarity Search, Recommender Systems, Visual Analytics

1 Introduction

Humans assess two objects as being similar if they are considered to be comparable with respect to certain properties. These properties can be either physical properties (e.g., dimensions, light reflectance, material, etc.) or semantic meta information (e.g., armchairs and chairs are functionally similar). For example, two books can be judged similar if they share similar content, or two movies if they have the same combination of genres. At the same time, two books can be similar because of equally colored covers and movies might be considered similar because of common actors. The notion of similarity is compound of different factors, including users, preferences, different options to define and measure properties, and also uncertainty. Besides the goal of searching for similar items, there are several other tasks that a user might want to accomplish. According to the *exploration-search axis*, introduced by Zahálka and Worring [28] in the field of Multimedia Analytics, there are two extreme values, namely Exploration and Search. In between those tasks, there are a variety of other tasks such as *Browsing, Summarization, and Ranking, etc.* which have to be considered as well when it comes to effective retrieval since an analytical work-flow may not only consist of similarity-search.

Digital data storage and processing enabled the research of automated similarity queries and founded the scientific area of information retrieval. A manual search for similar objects might be appropriate for small collections. However, with the advent of computers, the size of collections typically found is increasing rapidly. Prominent examples are *Spotify* with over 30 million songs and *Amazon* with over 200 million products. These volumes of information clarify the need for (semi-)automatic methods to retrieve and rank data items. A first mentioning of automatic retrieval of similar objects by Holmstrom [10] dates back as far as 1948. However, due to increasingly more complex objects, larger collections, and new user demands, automated similarity assessment is still an active research field. The existence of challenges, such as the Netflix Prize and conferences, such as the ACM RecSys, illustrate the practical importance and relevance of working on data- and user-adaptive similarity search. Among other things, the interaction with Recommender Systems (RSs) and helping users understand how their actions influence the recommendations are open challenges in the field of RSs [20]. The effectiveness of a RSs is dependent on more factors than just the quality of the similarity assessment method alone [26]. Similarity search should create trust, should be comprehensible, and transparent. In this paper, we identify interdepending factors influencing similarity search. We highlight arising research aspects and envision a Visual Analytics approach solving the introduced challenges.

2 Foundations

Many influencing factors need to be considered when engaging with the subject of similarity search. We categorize the influencing factors as *building blocks* of the respective *pillars* of similarity search. An overview about the identified pillars can be seen in Figure 1. In the following paragraphs, we describe and explain the three pillars *data*, *task*, and *user* in detail.

Data. Users need data to perform their retrieval tasks. Therefore, it is essential to pay particular attention when working with data, since errors made in early steps, for example during preprocessing, persist within the system and will negatively impact the quality of the results. In the case of IR or RSs, data might already be available beforehand, e.g., provided by a database with records of some kind of media (music, videos, products, images, etc.). Metadata describing the raw data, such as annotations, tags or derived data is usually available as well. Finally, there is also user-generated data. Bobadilla et al. [5]

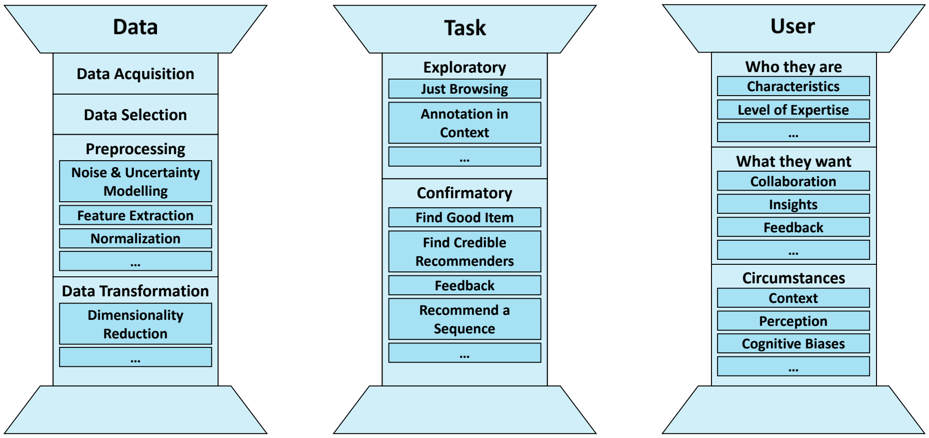

Fig. 1. An overview of the three pillars of similarity, *Data*, *Task*, and *User*. Each pillar consists of multiple building blocks, which in turn can have more building blocks.

describe the two ways of user data acquisition, in the case of IR and RS. Data can either be acquired explicitly, e.g., through user ratings, comments, etc., or implicitly, e.g., by the number of times a song was played. However, it is crucial to consider the noise or uncertainty in the data, especially for RSs, since there is not only natural, but also malicious noise [17]. Since RSs use real-world data, often provided by users, preprocessing is vital to enable similarity search providing relevant results. In data preprocessing, relevant descriptive features are derived and computed. These features should describe the represented objects very accurately and simultaneously enable similarity assessments. Amatriain et al. [3] give an overview of preprocessing methods in the context of RSs. The choice of the right similarity measure, for example, should be appropriate for the underlying data, even when already dealing with abstract representation of objects, such as feature vectors. The computation of feature vectors and the selection of similarity (distance) measures is highly domain dependent. A good example for the domain dependency are *tf-idf* vectors for the retrieval of text documents, where cosine similarity is the appropriate choice, since it ignores the length of the text documents and finds items of similar content. However, the similarity measure of genetic code – represented by letters and being textual data from an abstract point of view – employs other algorithms such as, Levenshtein, Needleman-Wunsch [16] or Smith-Waterman [25]. This holds also true for other types of data. Lew et al. [14] provide an overview of such data types.

Task. Tasks in similarity search have different backgrounds and goals, e.g., to explore data and formulate a hypothesis or to confirm/reject an existing hypothesis. Herlocker et al. [9] define eleven common tasks in which RSs are beneficial and helpful for users. We use this tasks exemplary to illustrate how the user's task influences the similarity search. The core recommendation task is to *Find Good Items* with respect to a specific information need. Early RSs [24] implemented this task by providing the user with a sorted list of results. For this kind of task, a range or k-nearest neighbor (k-NN) query using a classical similarity method is sufficient. However, depending on the definition of *good items*, adjustments of the similarity method are needed. For instance, the task *find good hairdressers* should not only consider the rating, but also the location, price, user preference, etc. Another important task is to *Find All Good Items*. In this case, neither the range nor k is known, hence a simple range or k-NN query is not sufficient for this kind of task. A simplified assumption would be, that all good items belong to the same cluster. Then instead of searching for the items themselves, one could search for the nearest cluster prototype. This task is especially important for lawyers or patent examiners, where missing one item can have a great impact. A third task is *Just Browsing*. Here, the user wants to explore the item or data space without a clear objective or information need. The similarity search should provide users with new items that might be of interest.

User. The user, applying similarity search to fulfill tasks on data, is judging the success or failure of the similarity-based application. User requirements are often complex and not always free of ambiguity. Users need to be considered not only by their ways of interaction but also by their characteristics and the search context. Users can have different levels of expertise in one or another field [18]. Behavioral scientists, for example, search for movement patterns differently than sport scientists might. Humans are intrigued by their own perspectives and insights. People are, consequently, often working collaboratively to satisfy their information need. Many more important characteristics for users exist and influence the perceived similarity such as a user's current location or time of day [1]. Additionally, not only the context, but also the perception and cognitive biases of the user have an influence during and after searching [13]. Currently, users are integrated into the process of similarity search by giving explicit feedback, for instance, by rating an item, or implicitly by analyzing the items a user previously viewed or for how long she or he viewed these items. Also, in E-commerce, metadata available on the users are exploited to learn and predict user preferences. Therefore, for a successful similarity search, it is key to understand who the users are, what they want to achieve, and under which circumstances they work with the similarity search.

3 Research Aspects

Although similarity search is a well researched and discussed area, there are many open challenges to tackle. Research aspects are categorized with respect to the previously introduced facets of similarity search. This Section is not intended to be a complete and exhaustive survey of the state-of-the-art, since this would exceed the scope of this paper. We rather envision and describe areas in which future research has to cope with open questions.

Data Accessibility and Usability. One constantly increasing main challenge for nowadays similarity search is that the employed data are often not accessible and usable enough. The *curse of dimensionality* [4], for example, falsifies the assumption that as more describing features are used, the similarity assessment will improve. Instead, severe effects on the similarity search have to be expected with an increasing number of dimensions. A dataset containing 15 dimensions, for example, can have a distance between the nearest neighbor close to the distance of the farthest neighbor. Although state-of-the-art similarity methods [15, 11] have shown that similarity search in high-dimensional data is possible to a certain extent, the selection of proper discriminative features and a semantic meaningful combination is crucial and complicated. Another challenge dealing with data is the preservation of privacy as stated in [8, 20]. Besides ethical and legal issues it is important to ensure that the intersection of query results of different data sources does not reveal more information than intended.

Models for Data and Context. On top of the data accessibility exists a noticeable lack when it comes to appropriate data and context models. This lack of data and context models is immediately affecting all of the introduced pillars in Section 2. For example, automatic methods cannot detect, handle, and remove all noise and uncertainty in the available data of RSs [17]. This can, for example, be illustrated by restaurant recommendations, assuming we have restaurants with noisy data of natural or malicious origin. Should a restaurant with a noisy rating still be considered as a *good item*, if it has otherwise positive attributes such as price and location? Furthermore, offering context-depending results of a similarity search helps in recommending *good items* [1, 8].

Visualization and Interaction. Eventually, the easiest way to provide a user with relevant items is to purely rely on the data and a static similarity measure. However, incorporating users by capturing their feedback and allowing them to modify the query and/or the similarity measure already improves the performance [23]. Nevertheless, visualizing an abstract similarity space and explain why results were found or not found is highly application and user dependent. Additionally, a lack of traceability combined with missing transparency [20] may lead to situations in which users are unaware where their insights came from and how the interactions with the system generated the results. As a consequence, the task might change during the process of analysis. *YouTube*, to name one famous example from one of the largest RSs in the world, uses Deep Neural Networks [6] for its recommendations. The shift towards a deep learning approach comes at the loss of transparency. For a given recommendation it is vague, how the data was weighted and which factors influenced the result. However, there are initial works proposing visualizations for neural networks that might help to overcome this problem. For instance, from Rauber et al. [19], which enables the inspection of relationships between neurons and classes.

4 Methodology

In Section 3 we described how the multitude facets of similarity search are influenced and influence each other. Understanding how these facets interdepend is crucial in order to improve the design of IR and RSs. In the following, we envision how such a system could be designed to support similarity search in the best possible way.

As we need various opportunities to reflect expert knowledge in the analysis process, we propose to follow a Visual Analytics process, as described by Keim et al. [12]. In Visual Analytics, heterogeneous data sources are processed and used to generate visualizations and models, thus enabling users to apply visual as well as automatic analysis methods. By interacting with the visualizations users are able to share background knowledge and context information via interactions. This information is then used to update the underlying model, which creates or updates models and visualizations. Following such a tight coupling of user and system will result in a continuous and mutual discourse, which will lead to higher confidence and better results.

A high-level description of the human and computer processes in Visual Analytics is given by Sacha et al. [22]. It helps to facilitate an understanding of the individual components and concepts of the Visual Analytics process and their interactions. Their *Knowledge Generation Model for Visual Analytics* can serve as a guideline on how to design new Visual Analytics systems or how to evaluate existing ones. One recent example where this is illustrated is the *Note Taking Environment* of Sacha et al. [21], which design is based on the knowledge generation model. Additionally, they show how Visual Analytics systems can be evaluated by measuring and investigating the trust of the user in the system.

In order to show how applying the Visual Analytics process can help tackle the open research aspects presented in Section 3, we incorporated them at the corresponding component in the Visual Analytics process, as illustrated in Figure 2. With the iterative and interconnected model for Visual Analytics, we are able to reflect the interdependent nature. This enables us to develop an understanding of the interdependencies of the different facets of similarity search and how Visual Analytics can help to tackle the open research aspects. The rationales behind this integration are outlined as follows.

Both the Visual Analytics model as well as our proposed pillars of similarity search have a data component, which serves as a base for the automatic analysis via data mining or similarity methods. With respect to the previously stated research aspects, data accessibility and usability questions are faced here. The transformation of the original raw data into meaningful and descriptive features is key for a successful similarity search. This transformation step is often also iterative and influenced by the curse of dimensionality, especially in the design phase of a similarity search Visual Analytics system.

Models for data and context influencing the similarity search as described as the second research aspect are key to understand how users employ RSs. Another important aspect which still needs more attention in the field of RSs are visualizations of both, the results and the underlying model [8]. It is not

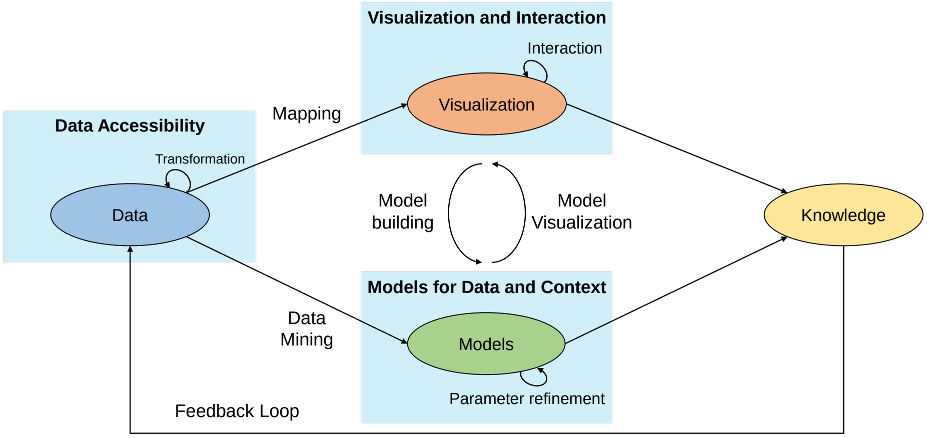

Fig. 2. Main research questions in similarity search integrated in the Visual Analytics model of Keim et al. [12]. The iterative and interconnected model of Visual Analytics reflects the interdependencies of our described research challenges.

only important to identify, how a user can interact with these visualization [2], but also what the rationales behind these interactions are [27]. By capturing these interactions, as well as contextual information, conclusions about the goals of the users can be drawn. This enables us to train the underlying model of the similarity search according to their expectations. Consequently, by Visual Analytics we are able to enrich the similarity search by “Insight Provenance” and traceability of the results.

As the key ingredients of Visual Analytics are visualization and interaction, the overlap to the third research aspect is granted per se. Visualization and user interaction can be used to utilize the user’s domain knowledge [7]. In a two-dimensional spatial visualization of documents, documents are distributed by their similarity to each other. By spatially rearranging documents, for example by drag-and-drop, users can communicate to the system, which documents they find similar, which in turn trains the similarity model according to their feedback. As a consequence, the user’s domain knowledge is captured, interpreted, and applied to the whole dataset.

5 Conclusion

We believe following a Visual Analytics approach will improve similarity search applications, in particular IR and RSs. With the user-centered focus of Visual Analytics combined with data analytics, information visualization, and interaction, query results can be made transparent and interpretable. Finally, transparent query results will increase users’ trust in the similarity search results. However, as a direct consequence of applying the Visual Analytics process on similarity search, new challenges are emerging. There is a need for an increased understanding of the relationship of the components in the process and the influences of the various parameters. This can lead to new insights which help to identify errors, improve robustness, and increase quality of, as well as trust in similarity search.

References

1. Adomavicius, G., and Tuzhilin, A. Context-aware recommender systems. In Recommender systems handbook. Springer, 2015, pp. 191–226.
2. Amar, R., Eagan, J., and Stasko, J. Low-level components of analytic activity in information visualization. In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on (2005), IEEE, pp. 111–117.
3. Amatriain, X., Jaimes, A., Oliver, N., and Pujol, J. M. Data mining methods for recommender systems. In Recommender Systems Handbook. Springer, 2011, pp. 39–71.
4. Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is nearest neighbor meaningful? In International conference on database theory (1999), Springer, pp. 217–235.
5. Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. Recommender systems survey. Knowledge-based systems 46 (2013), 109–132.
6. Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems (2016), ACM, pp. 191–198.
7. Endert, A., Fox, S., Maiti, D., and North, C. The semantics of clustering: analysis of user-generated spatializations of text documents. In Proceedings of the International Working Conference on Advanced Visual Interfaces (2012), ACM, pp. 555–562.
8. He, C., Parra, D., and Verbert, K. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. Expert Systems with Applications 56 (2016), 9 – 27.
9. Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) 22, 1 (2004), 5–53.
10. Holmstrom, J. E. Section iii. opening plenary session. In The Royal Society Scientific Information Conference (1948), Royal Society.
11. Houle, M. E., and Sakuma, J. Fast approximate similarity search in extremely high-dimensional data sets. In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on (2005), IEEE, pp. 619–630.
12. Keim, D., Kohlhammer, J., Ellis, G., and Mansmann, F. Mastering the information age solving problems with visual analytics. Eurographics Association, 2010.
13. Lau, A. Y., and Coiera, E. W. Do people experience cognitive biases while searching for information? Journal of the American Medical Informatics Association 14, 5 (2007), 599–608.
14. Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2, 1 (2006), 1–19.
15. Liu, K., Bellet, A., and Sha, F. Similarity learning for high-dimensional sparse data. In AISTATS (2015).

16. Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48, 3 (1970), 443–453.

17. O'Mahony, M. P., Hurley, N. J., and Silvestre, G. Detecting noise in recommender system databases. In Proceedings of the 11th international conference on Intelligent user interfaces (2006), ACM, pp. 109–115.

18. Picault, J., Ribiere, M., Bonnefoy, D., and Mercer, K. How to get the recommender out of the lab? In Recommender Systems Handbook. Springer, 2011, pp. 333–365.

19. Rauber, P. E., Fadel, S. G., Falcao, A. X., and Telea, A. C. Visualizing the hidden activity of artificial neural networks. IEEE Transactions on Visualization and Computer Graphics 23, 1 (2017), 101–110.

20. Ricci, F., Rokach, L., and Shapira, B. Recommender systems: Introduction and challenges. In Recommender Systems Handbook. Springer, 2015, pp. 1–34.

21. Sacha, D., Boesecke, I., Fuchs, J., and Keim, D. A. Analytic behavior and trust building in visual analytics. In Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers (2016), Eurographics Association, pp. 143–147.

22. Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. A. Knowledge generation model for visual analytics. IEEE transactions on visualization and computer graphics 20, 12 (2014), 1604–1613.

23. Seebacher, D., Stein, M., Janetzko, H., and Keim, D. A. Patent Retrieval: A Multi-Modal Visual Analytics Approach. In EuroVis Workshop on Visual Analytics (EuroVA) (2016), N. Andrienko and M. Sedlmair, Eds., The Eurographics Association, pp. 013–017.

24. Shardanand, U., and Maes, P. Social information filtering: algorithms for automating word of mouth. In Proceedings of the SIGCHI conference on Human factors in computing systems (1995), ACM Press/Addison-Wesley Publishing Co., pp. 210–217.

25. Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. Journal of molecular biology 147, 1 (1981), 195–197.

26. Swearingen, K., and Sinha, R. Beyond algorithms: An hci perspective on recommender systems. In ACM SIGIR 2001 Workshop on Recommender Systems (2001), vol. 13, Citeseer, pp. 1–11.

27. Yi, J. S., ah Kang, Y., and Stasko, J. Toward a deeper understanding of the role of interaction in information visualization. IEEE transactions on visualization and computer graphics 13, 6 (2007), 1224–1231.

28. Zahálka, J., and Worring, M. Towards interactive, intelligent, and integrated multimedia analytics. In Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on (2014), IEEE, pp. 3–12.