Fostering Serendipitous Knowledge Discovery using an Adaptive Multigraph-based Faceted Browser

Ali Khalili
Department of Computer Science
Vrije Universiteit Amsterdam
a.khalili@vu.nl

Pek van Andel
The University Medical Center Groningen
m.v.van.andel@umcg.nl

Peter van den Besselaar
Department of Organization Sciences
Vrije Universiteit Amsterdm
p.a.a.vanden.besselaar@vu.nl

Klaas Andries de Graaf
Department of Computer Science
Vrije Universiteit Amsterdam
ka.de.graaf@vu.nl

ABSTRACT

Serendipity, the art of making an unsought finding plays also an important role in the emerging field of data science, allowing the discovery of interesting and valuable facts not initially sought for. Previous research has extracted many serendipity-fostering patterns applicable to digital data-driven systems. Linked Open Data (LOD) on the Web which is powered by the Follow-Your-Nose effect, provides already a rich source for serendipity. The serendipity most often takes place when browsing data. Therefore, flexible and intuitive browsing user interfaces which support serendipity triggers such as enigmas, anomalies and novelties, can increase the likelihood of serendipity on LOD. In this work, we propose a set of serendipity-fostering design features supported by an adaptive multigraph-based faceted browsing interface to catalyze serendipity on Semantic Web and LOD environments.

1 INTRODUCTION

"Unless you expect the unexpected you will never find [truth], for it is hard to discover and hard to attain." -*Heraclitus*[1]

The experience of 'accidental' discovery and acquisition of information generally known as *Serendipity* refers to 'accidentally' bumping into (new, true, useful, or personal interest-related) information, initially not looked for. Serendipity, defined as the art of making an unsought finding[24], has played a pivotal role in the discovery of many drugs. Major types of psychotropic drugs (effecting mental activity and behavior) such as *Lithium*, *Chlorpromazine* and *Imipramine* were serendipitously discovered in the 1950s and 1960s. In 2012, [7] reported that 24% of all pharmaceuticals on the market and in particular 35.2% of all the anticancer drugs in clinical use were discovered by serendipity.

Serendipity also plays an important role in the emerging field of data science by enhancing information retrieval[8] and by promoting unexpected knowledge discovery. The World Wide Web has provided a global information space comprising billions of connected documents. "The unexpected connection is more powerful than one that is obvious", as aptly asserted by Haraclitus in 500 BC. However, most of the existing centralized "nearest neighbor" search approaches on the Web, such as Google, although very useful in finding explicitly relevant results, are killing serendipity by excessively limiting the encountering of unexpected information[1]. On the other hand, the ever-growing amount of Linked Data publicly accessible and distributed on the Web increases the likelihood that some of the data, which will make an impact in our professional or private lives will come to us by chance—without searching it initially. The adoption of Semantic Web as a linked information space in which data are dynamically enriched and added, provides an open interactive system, with external links and the ability to make information easily accessible, re-usable including the possibility of the discovery and serendipitous reuse of other related information[2, 20].

'Unsought discoveries' most often take place in the context of browsing unbounded data spaces; people immerse themselves in the items that interest them, meandering from topic to topic, and so on and so forth (i.e., the *Follow-Your-Nose* method[26] to traverse the given semantic links from a resource) while concurrently remarking interesting and informative information en route[23]. Therefore, flexible and intuitive browsing user interfaces (UIs) which support serendipity triggers, can increase the likelihood of accidental knowledge discovery on Linked Open Data (LOD). Although there has been some research on supporting serendipity through query modifications and semantic path-finding on knowledge graphs, we still lack UIs that increase the emergence of serendipities on LOD. In this paper, we aim to provide an adaptive multigraph-based faceted browsing interface to foster serendipity on Semantic Web and LOD environments. The contributions of this work are in particular: 1) Proposing a set of serendipity-fostering design features which are applicable to data-driven environments, by conducting an extensive literature review. 2) Implementing an open-source adaptive multigraph-based faceted browser to support the proposed serendipity design features while browsing linked data.

2 SERENDIPITY: THE ART OF UNSOUGHT FINDING

The word "serendipity" was coined in 1754 by Horace Walpole, a letter writer and politician[15]. Walpole was inspired by an old

[1] according to secondary sources

K-CAP, 2017, US

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Persian fairy tale known as "The Three Princes of Serendip" published in 1302 AD by Amir Khusrow Dehlavi[2]. The original story is about three princes from Serendip (a medieval Persian name for Sri Lanka), well trained in the art of tracking, who make ten 'accidental' discoveries via ten surprising observations, and by interpreting all ten correctly, on their grand tour to see the different countries and miracles of the world. Walpole created the word serendipity to refer to "always making discoveries, by accidents and sagacity, of things they (the three princes of Serendip) were not in quest of" or "a surprising observation followed by a correct hypothesis".

In our view, serendipity consists of two main steps: a surprising observation (*trigger*) and then a correct interpretation (*abduction*). The trigger is a riddle, an anomaly, or a novelty. Abduction[25] refers to the process of guessing, interpreting, creating and testing hypotheses in order to find a correct explanation, one that is evidence-based. As stated in [18], you do not reach Serendip by plotting a course for it. You have to set out in good faith for elsewhere and lose your bearings...serendipitously!

Serendipitous discovery may be facilitated but it is by definition an emergent process[6]. Because it is an emergent process, transitioning serendipity to a science where certain patterns are defined is an inherently difficult task to be managed. What we can offer is to foster the process of serendipity by providing an incubator-like environment for serendipity. In other words, the environment will increase the likelihood of serendipity, without guaranteeing it. In the context of a knowledge building environment, [1] calls such a system that supports both *trigger* and *abduction* an "inspiration engine" or in [7], the term "pseudo-serendipity" is used to describe the approach of such systems i.e. *a sought finding, found on an unsought road*. As result of our extensive literature review and consultation with a serendipitiologist, we extracted, blended and adapted a set of serendipity-fostering design features that are applicable to data-driven systems. The main sources of inspiration for these features come from [3, 5, 23, 24]. These features can coexist, overlap, cooperate, complement and reinforce each other. In the following sections, we describe these 12 serendipity-fostering design features together with some ideas how to realize them:

2.1 Design Features Related to Observations

F_1: Make surprising observations more noticeable.

Surprising observations are the main initiators of accidental knowledge discovery. A single surprising observation, especially if it is repeatedly done, or multiple different surprising observations, when they refer to the same phenomenon, can trigger serendipity. Creative data visualization is an activity that enables users to make hypotheses, look for patterns and exceptions, and then refine their hypothesis. Users might find surprising results that shake their established beliefs, provoke new insights, and possibly lead to important discoveries[21]. Users often need to look at the same data from different perspectives. Therefore, tools that provide different views on data can foster serendipity.

F_2: Make errors in data more visible in order to detect successful errors easier.

Errors and exceptions are not always accidental and can sometimes indicate the real and natural behavior of a system known as "desire lines"[16]. Following the trails left behind quantitative and qualitative anomalies in data can result in new insights. Semantic Web tools, such as Shapes Constraint Language (SHACL)[3], or restrictions supported by RDF-S & Web Ontology Language (OWL), which allow validating RDF graphs against a set of rules and conditions, help to automate the discovery of successful errors and thereby facilitate serendipity.

F_3: Allow inversion and contrast.

The inversion and contrast features depict the unexpected aspect of serendipity. Sometimes turning things upside down or inside out allows us to watch those things from another perspective and to discover gaps in knowledge. Looking at the insights in the opposite direction than intended by users will and can cause to a breakthrough discovery. This feature can be supported by SPARQL query inversion where a query is adapted to include results which were not returned by the initial query; or the query employs RDF properties which contrast with the initial properties used.

F_4: Support randomization and disturbance.

Chance can be used intentionally in serendipitous knowledge discovery. Randomization and disturbance are two methods to increase the chance encounter. Randomized Coffee Trial (RCT), is a technique used by some firms to create an institutionalized space for serendipity through connecting people in the firm at random and give them time to meet to have a coffee and talk about whatever they wish. In a linked data browsing system, randomizing the items (or modifying the order of the sets of triples) presented on top of the result lists can increase the probability of the chance encounter. It also serves as an efficient solution to the problem of 'blind spots' and to decrease the possibility of bias in interpreting results.

F_5: Allow monitoring of side-effects when interacting with data.

Accidental discoveries through observation of side-effects has played a crucial role in drug/treatment discovery. For example, *Dimenhydrinate* was first developed as an antihistamine, but is now sold as a travel sickness medication owing to a surprising observation/realization by one of the participants in the clinical trials[7]. A system that consistently monitors the side-effects of user interactions with data and provides appropriate feedback on surprising observations implied, can facilitate serendipity.

F_6: Support detection and investigation of by-products.

Some serendipitous discoveries have occurred as by-products or spin-offs of the main product which was intended to come out. A user searches for A and, as a by-product, finds B as a surprising unsought result. A system that supports detection and investigation of by-products resulted from user interactions can foster serendipity. Error, enigma, anomaly and novelty detection mechanisms suggested by F_4 can support this feature as well.

F_7: Support background knowledge and user contextualization.

"In the sciences of observation, chance favors only prepared minds", said Louis Pasteur. Most serendipitous discoveries are triggered by chance or a chance encounter. A chance encounter occurs at the point in human interaction with an information system when a human makes an accidental discovery. The encounter is generally influenced by the person's prior knowledge, although not necessarily, and by the person's recognition of the affordances. A serendipity-fostering environment depends both on the information seeker

[2] https://en.wikipedia.org/wiki/Amir_Khusrow

[3] https://www.w3.org/TR/shacl

and the medium. Without basic topical knowledge, there is no capacity to observe and interpret the surprising facts correctly[8]. Techniques for integrating user profiles and domain knowledge into query processing[22] can improve the relevancy of the query results obtained by users and thereby promote the serendipity.

F_8: Support both convergent and divergent information behavior. When users move through an information space they may change directions and behavior several times as their information needs and interests develop or get triggered depending on affordances encountered on their way through the information space. Supporting both convergent and divergent information behavior[3] in a data-driven system facilitates serendipity. Convergent (depth first, focused, not easily distracted) behavior is supported by features that allow zooming in and narrowing the vision of users while divergent (breadth first, creative, but easily distractible) behavior is supported by features that allow zooming out and broadening the vision of users.

2.2 Design Features Related to Explanation of the Observations

F_9: Facilitate the explanation of surprising observations. After the occurrence of a surprising observation as trigger, abduction is needed to understand why and how this accidental event is entailed. Abduction gives some clues to interpret the surprising result and to find the correct explanation for it. Metadata and provenance as means to support causal reasoning aid to provide reasons and explanation for surprising observations. Provenance also helps to assess the quality, reliability, or trustworthiness of surprising data which is discovered. Exploiting existing provenance data models and ontologies[4] on the Semantic Web can foster serendipity.

F_{10}: Allow sharing of surprising observations among multiple users. A surprising observation done by user A, when correctly explained by user B, can result in positive serendipity. Tools such as *YASGUI*[5], *grlc*[6] and *BASIL*[7] help to support this feature via sharing and modification of SPARQL queries among multiple users through a standardized Web API.

F_{11}: Enable reasoning by analogy. Analogical learning as the act of finding similar entities or phenomena when studying an entity or phenomenon has been long known as an approach for knowledge transfer. Analogical reasoning can happen either in the same or a completely different context than the original context of data. Semantic Web-based knowledge abstraction techniques on LOD help to foster serendipity by enabling the abstraction of the knowledge representation structure related to a particular knowledge artifact, by analyzing its constituent elements and their relationships. For instance, by employing SPARQL query patterns one can identify similar resources to a resource of interest by considering resources with similar RDF properties and values or with more generalized RDF classes than the resource of interest. With regards to analogical reasoning on different context (e.g. concepts from business domain which are similar to concepts in medical domain), there are less strategies discussed in the literature.

[4] https://www.w3.org/TR/prov-overview
[5] http://yasgui.org
[6] http://grlc.io
[7] http://basil.kmi.open.ac.uk

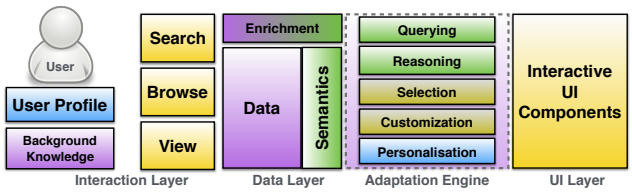

Figure 1: Architecture of the proposed adaptive faceted browsing environment.

In [13], a framework is proposed for explicitly modeling analogical structures in multi-relational or knowledge graph embedding. Another possible strategy is to analyze ontology design patterns instead of concrete entity-similarity metrics to represent relevance between entities in one context to entities in another context[4].

F_{12}: Support extending the memory of user by invoking provocative reminders and relevance feedback. Keeping track of previous user interactions, queries, and resulting data while browsing complex data enables a data-driven system to invoke provocative messages as reminders to help extend the memory of users when interacting with other related datasets. When potentially valuable information is encountered an important ability would be the capacity to recognize it and its "affordances"[17]–clues about how it can be used. *Relevance feedback*[14]–asking information seekers to make relevance judgments about returned objects and then executing a revised query based on those judgments–is already known as a powerful way to cultivate knowledge discovery. If a person is not alert enough, the message remains unnoticed regardless of its potential value. A system that provides users with meaningful reminders connected to their past browsing experience can increase the likelihood of serendipity.

3 AN ADAPTIVE MULTIGRAPH-BASED FACETED BROWSER: A TRIGGER AND FACILITATOR FOR SERENDIPITY

There are generally three ways in which people discover and acquire information: 1) The *Purposive search*: A directed search looking for a definite piece of information. 2) *Exploratory search and browsing*: A general purpose semi-directed search and browsing of data deliberately looking for an object that cannot be fully described or to get inspiration by looking at some items of interest. 3) *Capricious search and browsing*: An undirected random search and browsing of information without a defined goal. Accidental knowledge discoveries occur most frequently during this type of unplanned investigative search and browsing.

Systems that support the first, prompt users for search terms and keywords, and provide options for *parametric search* allowing users to manipulate queries and results by visually specifying a set of constraints. The focus of such systems which are well supported by current Web search engines is on *precision* i.e. minimizing the number of possibly irrelevant objects that are retrieved.

Hypermedia, menu-driven and faceted navigation systems that provide views and overviews of the data facilitate the second. Faceted navigation fills in the piece that is missing in parametric search: *guidance*. Parametric search requires that the user express an information need as a query in one shot, making selections

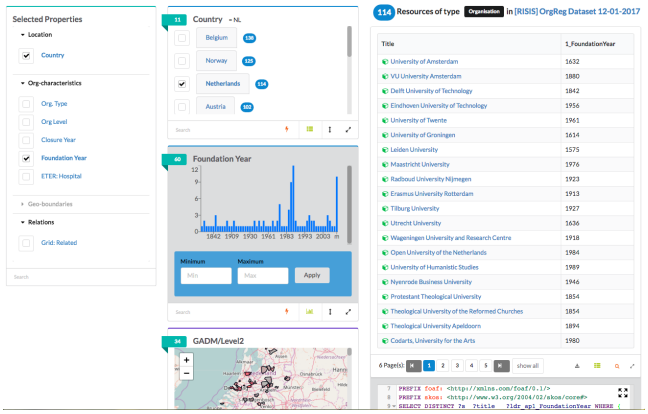

Figure 2: An screenshot of the implemented faceted browser.

across all facets of interest. In contrast, *faceted navigation* allows the user to elaborate a query progressively, seeing the effect of each choice in one facet on the available choices in other facets. Faceted browsers are often seen as most promising candidates for rich exploration of a domain across a variety of sources from a user-determined perspective [19]. Systems that support this type of browsing are more concerned with recall i.e. maximizing the number of possibly relevant objects that are retrieved.

The third type, the serendipitous approach, is a type of information seeking that is not traditionally examined in information retrieval research and has received little attention by both developers and researchers. In this paper we focus on this latter type by augmenting the existing faceted browsing techniques with serendipity-fostering features discussed in Section 2. We call our proposed adaptive faceted browsing environment "FERASAT"[8] (FacEted bRowser And Serendipity cATalyzer). FERASAT is built on top of the LD-R framework[11] to enable skeuomorphic, adaptive and component-based design of the system. Skeuomorphism[17] in UI design is employed to incorporate recognizable UI elements which are familiar to users and thereby decrease the cognitive load of users when interacting with the system. Skeuomorphic design in FERASAT is a way to bypass the *Pathetic Fallacy of RDF*[9][9].

Figure 1 depicts the architecture of the FERASAT where related elements are color coded. The system provides three main modes of interaction with data namely search, browse and view. During the user interactions, based on the semantics of data and the given user context, the system adapts its behavior by rendering appropriate interactive UI components. FERASAT provides a particular type of adaptive UI called a *data-aware UI* [10] that a) can understand users' data and b) can interact with users accordingly.

FERASAT is implemented as an open-source project which is available to download at http://ferasat.ld-r.org. (see Figure 2 for an screenshot of the FERASAT environment). In [12], we provide an extended version of this paper which consists of a more detailed technical description of the system together with the discussion of the related work and a set of use cases.

[8] in Persian, the term 'Ferasat' refers to the ability of intuitive knowledge acquisition.
[9] i.e. display RDF data to the users as a graph because the underlying data model is a graph.

4 CONCLUSIONS

Linked Open Data provides a rich domain for people to experience serendipity — finding valuable or agreeable things initially not sought for. Serendipity is a by-catch, an outcome or a moment of successful retrieval when a user is browsing data. In this paper we presented a set of serendipity-fostering design features amenable to data-driven systems together with a set of UI and Semantic Web techniques which support those features when a user is exploring linked data on a faceted browsing environment. To showcase the applicability of our proposal, we implemented a data-aware faceted browser UI to foster accidental knowledge discovery while browsing data scattered over multiple knowledge graphs.

REFERENCES

[1] A. Acosta. Using serendipity to advance knowledge building activities. *Ontario Institute for Studies in Education, University of Toronto, Canada*, 2012.
[2] G. Alemu, B. Stevens, P. Ross, and J. Chandler. Linked data for libraries: Benefits of a conceptual shift from library-specific record structures to rdf-based data models. *New Library World*, 113(11/12):549–570, 2012.
[3] L. Björneborn. Design dimensions enabling divergent behaviour across physical, digital, and social library interfaces. In *PERSUASIVE*. Springer, 2010.
[4] F. Bobillo, M. Delgado, and J. Gómez-Romero. An ontology design pattern for representing relevance in owl. *The Semantic Web*, pages 72–85, 2007.
[5] P. H. Cleverley and S. Burnett. Retrieving haystacks: a data driven information needs model for faceted search. *Journal of Information Science*, 41(1), 2015.
[6] M. Cunha. Serendipity: Why some organizations are luckier than others. *Universidade Nova de Lisboa (Ed.), FEUNL Working Paper Series*, 2005.
[7] E. Hargrave-Thomas, B. Yu, and J. Reynisson. The effect of serendipity in drug discovery and development. *Chemistry in New Zealand*, 2012.
[8] J. Heinström. Psychological factors behind incidental information acquisition. *Library & Information Science Research*, 28(4):579–594, 2007.
[9] D. Karger and M. Schraefel. The pathetic fallacy of rdf. SWUI, 2006.
[10] A. Khalili and K. A. de Graaf. Linked data reactor: Towards data-aware user interfaces. In *Proceedings of the 13th International Conference on Semantic Systems*, SEMANTiCS 2017. ACM, 2017.
[11] A. Khalili, A. Loizou, and F. van Harmelen. Adaptive linked data-driven web components: Building flexible and reusable semantic web interfaces. In *ESWC2016*, pages 677–692, 2016.
[12] A. Khalili, P. van Andel, P. van den Besselaar, and K. A. de Graaf. The three princess of serendip on linked data, 2017. http://research.ld-r.org/papers/three-pincess-LinkedData.pdf.
[13] H. Liu, Y. Wu, and YimingYang. Analogical inference for multi-relational embeddings. https://arxiv.org/abs/1705.02426, 2017.
[14] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, Apr. 2006.
[15] R. Merton and E. Barber. *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*. Princeton Press, 2004.
[16] D. A. Norman. *Living with complexity*. MIT press, 2010.
[17] D. A. Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, Inc., New York, NY, USA, 2013.
[18] N. Ramakrishnan and A. Y. Grama. Data mining: From serendipity to science. *Computer*, 32(8):34–37, 1999.
[19] G. M. Sacco and Y. Tzitzikas. *Dynamic taxonomies and faceted search: theory, practice, and experience*, volume 25. Springer Science & Business Media, 2009.
[20] N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101, 2006.
[21] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *AVI2006 workshop*, pages 1–7. ACM, 2006.
[22] V. C. Storey, V. Sugumaran, and A. Burton-Jones. The role of user profiles in context-aware query processing for the semantic web. In *Intl. Conf. on Application of Natural Language to Information Systems*, pages 51–63, 2004.
[23] E. G. Toms et al. Serendipitous information retrieval. In *DELOS Workshop*, pages 17–20, 2000.
[24] P. van Andel. Anatomy of the unsought finding. serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *Br J Philos Sci*, 45(2):631–648, June 1994.
[25] P. van Andel and D. Bourcier. Serendipity & abduction in proofs, presumptions & emerging laws. In *The Dynamics of Judicial Proof*, pages 273–286. 2002.
[26] L. Yu. *Follow Your Nose: A Basic Semantic Web Agent*, pages 711–736. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.