Towards Integrity in Diversity-aware Small Set Selection and Visualisation Tasks

Marcin Sydow[1,2]
[1]*Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*
[2]*Polish-Japanese Institute of IT, Warsaw, Poland*

Keywords: Diversity, Integrity, Summarisation, Semantic Graphs, Visualisation.

Abstract: In this short paper, we introduce a novel notion of *integrity* in diversity-aware selection and visualisation tasks, present motivation for studying this notion and illustrate it on a case study concerning the visualisation of semantic entity summaries. In particular, we propose a novel visual integrity measure for this case study and illustrate it in a preliminary experiment.

1 INTRODUCTION

Diversity is a desired property of results returned to the user in many applications concerning selecting a small set of information pieces. It is especially true in cases when the actual user information need is unknown or ambiguous and the limit on the amount of presented information is low. This concerns numerous important practical tasks ranging from result diversification in Web search (Agrawal et al., 2009), database querying (Vee et al., 2008), recommender systems to diversity-aware text summarisation (Carbonell and Goldstein, 1998) and recently entity summarisation (Xu et al., 2014) and more specifically *semantic entity summarisation* (Xu et al., 2014; Sydow et al., 2013), to mention some examples. The main idea in such diversity-aware approaches is to select and present to the user pieces of information that are not only potentially maximally *relevant* to the user's information need but also maximally *diversified*. This is achieved by various techniques. For example, some of them introduce a pair-wise dissimilarity measure between the presented items and approach it as an adaptation of the Maximum Facility Dispersion Problem (Gollapudi and Sharma, 2009). Other view the problem as the maximum coverage (Clarke et al., 2008) or as minimising the probability of query abandonment (Chen and Karger, 2006). The rationale behind such approaches is to avoid redundancy in presented information and the optimal use of the given low limit on the amount of presented information i.e. to cover maximally many different possible aspects/interpretations of the presented information to satisfy at least partially an unkown actual user's information need.

1.1 The Issue of Information Integrity

The information selection tasks mentioned above can be roughly divided into two cases.

First Case: Independent Items. In the first group, each separate returned item from the result set represents rather an independent piece or portion of useful information to the user itself. This concerns: each returned link to a web document in web search result set, each returned record in database query result set or each separate recommended item in the result set returned by a recommender system.

In such cases, increasing the diversity of results, *without* loosing relevance generally improves the quality of results.

Second Case: Inter-Dependent Items. There is a second group of tasks, however, that should be treated in a different way. Here, the returned items can be more inter-dependent. This concerns especially all summarisation-like tasks. For example, in extractive text summarisation, the task is to select a small set of sentences out of the input text that summarise it. Notice that a single piece of useful information can be spread among two or more sentences that refer to each other, in such case a single sentence does not necessarily represent a full sense to the user, alone. Similar situation concerns the *entity summarisation* task (Xu et al., 2014; Sydow et al., 2013), where given an input entity, the output result is expected to be a representative set of facts or features selected from an underly-

Sydow M..
Towards Integrity in Diversity-aware Small Set Selection and Visualisation Tasks.
DOI: 10.5220/0005159904800484
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 480-484
ISBN: 978-989-758-048-2
Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

ing knowledge base and presented to the user. In such case some facts or features concerning the entity make more sense to the user only when presented *together* in the *context of* the summarised entity. For example, the properties such as "longitude" and "latitude" in the context of an entity that has some geographical location (e.g. a city) make a full sense to the user only when presented together. We will hence refer to the issue described in such case as to "information integrity".

1.2 Motivation and Contributions

It is important to observe that in most of the diversity-aware approaches known from literature such a information link between two or more inter-dependent items will be ignored (in optimistic scenario) while it should not as it may lead to separating the pieces of information. Furthermore, in the diversification techniques that are based on redundancy-avoiding such items will be very likely to be identified as "similar" and consequently *separated* to avoid redundance in the result. For example, when item "latitude" is selected to be presented in the entity summary, the item "longitude" will be dropped as being "similar". As the result the user would not obtain complete information.

The main motivation of this paper is to introduce the notion of "information integrity" to the research concerning diversity-aware approaches and to propose some fundamental observations and concepts in order to start the discussion on this issue.

1.3 Integrity-awareness in Three Phases

In particular, we identify three basic phases of a diversity-aware information interaction process, where the information integrity issue should be properly addressed:

1. pre-selection phase: automatic detection of the "information integrity" issue (e.g. computing some statistics to identify pairs or sets of pieces of information that should be treated as "integral")
2. selection phase: enriching the diversity-aware result selection phase with integrity-awareness (e.g. to avoid separation of ensembles of items that have been identified as "integral" and presented together)
3. presentation phase: enriching the presentation of the results with integrity-awareness (e.g. to emphasise in the presented results the integrity of some subsets of presented items)

We realise that each of the above points deserves for a separate study and concrete solutions can strongly depend on the particular application.

We hope that the proposed ideas are to some extent adaptable in other applications, in particular in other information visualisation tasks (for example Google Knowledge Graph or Semantic Knowledge Graph browsers (e.g. Yago Browser[1])

2 CASE STUDY: INTEGRITY-AWARE VISUALISATION IN THE DIVERSUM PROBLEM

In this section we illustrate integrity-awareness problem on an example. The example is a specific summarisation problem called DIVERSUM, presented quite recently in (Sydow et al., 2013). The full name of the DIVERSUM problem is: diversity-aware entity summarisation on semantic knowlege graphs. The specification of the problem will be given in Section 2.2. More precisely, to illustrate the integrity-awareness issues we focus on the visualisation phase of the DIVERSUM problem.

2.1 Data: Semantic Knowledge Graph

In this problem, there is given an underlying semantic knowledge graph KG. In short, it consists of a large collection of so-called subject-predicate-object triples, where subject and object are some nodes in KG. The nodes can represent entities – in such case the predicate represents a "fact" concerning two entities (e.g. (Woody Allen, acted in, Zelig). One or both nodes can also represent classes, e.g. (Woody Allen, has type, Actor) or (Actor, is subclass of, Person), etc.

2.2 DIVERSUM: Problem Specification

The problem has the following specification:[2]
INPUT:

1. KG – the underlying knowledge base
2. q – a node of KG (entity to be summarised)
3. $k \in N$ – an upper limit on the number of facts (triples) to be presented in the entity summary

OUTPUT: S: summary of entity q – a connected subgraph of D containing q and at most k arcs that together represent a collection of facts being a summary

[1] https://gate.d5.mpi-inf.mpg.de/webyagospotlx/SvgBrowser

[2] For a fuller discussion of the DIVERSUM problem we refer the reader to (Sydow et al., 2013)

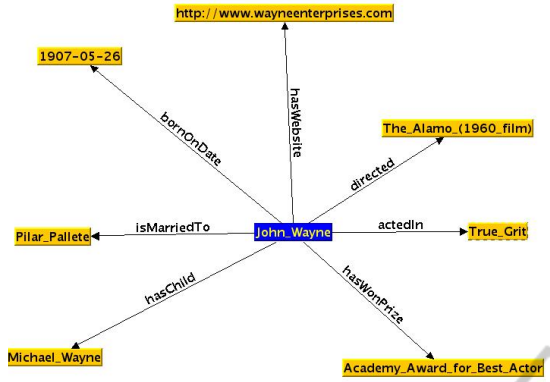

Figure 1: Example graphical entity summary of John Wayne computed on imdb knowledge graph by DIVERSUM algorithm with k=7.

of information concerning the entity in the semantic knowledge graph. In the DIVERSUM problem we additionally pay attention to make the summary diversified.

An example concerning imdb movie semantic database[3] of the resulting summary in the DIVERSUM problem is presented on Figure 1. The entities in this dataset represent movies, actors, prizes, etc. We will call each particular triple as "fact" (e.g. "John Wayne, acted in, True Grit") and each arc label as "predicate" (e.g. "acted in"). Using the general terminology as in Section 1 the *items* to be returned are the facts concerning the input entity.

In the DIVERSUM problem the result is *graphical*. Thus, the problem does not consist only in *selecting* the items to be shown in the entity summary(selection phase mentioned in Section 1.3) but also in a subsequent visualisation of the results (presentation phase mentioned in Section 1.3). In the remaining part of this paper we will illustrate how the integrity-awareness issue can be naturally observed in the visualisation phase of the DIVERSUM problem.

2.3 Integrity in Visualisation Phase of DIVERSUM

When the set of facts to be presented is selected the remaining task is to decide the *layout* of the summary. We will simplify the problem and focus only on the *order* of the facts to be presented. Notice that in the DIVERSUM problem, the presented facts concerning the summarised entity form a *ring* of k facts (Fig. 1).

[3]www.imdb.org

2.4 Integrity-aware Visualisation as an Optimisation Problem

We propose that in the visualisation phase the goal of the integrity-aware approach can be simply expressed as follows: *similar elements should be shown close to each other*.

We propose to specify this problem as an optimisation one. More precisely, it can be done by defining *integrity measure* of a given layout and select the layout that maximises the integrity measure.

2.5 Proposed Visual Integrity Measure

A visual integrity measure should promote layouts in which similar elements are close to each other. Let assume that S is the $k-$set of selected items to be presented in the summary. Let's also assume that there is defined a pair-wise semantic similarity measure $sim : S^2 \to Q+$. We assume that sim is symmetric (i.e. $sim(a,b) = sim(b,a)$). The higher the value, the more similar are the items.

We propose to consider the following visual integrity measure *vim to be maximised* based on sim:

$$vim(L) = \sum_{s \in S} sim(s, next_L(s))$$

where L denotes particular layout of the selected items and $next_L(s)$ denotes the element $s' \in S$ that is shown *next* (say, clockwise)[4] to s in the layout L. The interpretation of vim is simple: it sums up similarity measures of *neighbouring* elements in the layout. By maximising vim we force the layout to show similar elements next to each other, in a way.

2.6 Experimental Example

To illustrate the discussed ideas we will use the *imdb* movie dataset converted to the format of a semantic knowledge graph.

In this example, for illustration, we define the underlying pair-wise similarity sim measure in a simple way, based on co-occurrence statistics in the imdb dataset. More precisely, for two facts a,b, the similarity measure value $sim(a,b)$ is defined as the number of entities in the dataset D in which the predicate names in a,b co-occur divided by the number of entities in D in which at least one predicate name from a,b occurs. In other ways it is an adaptation of the Jaccard co-efficient. To explain: assume that there are 200 entities incident with predicates $a =$"actedIn"

[4]This choice is arbitrary, counter-clockwise would result in the same value in this case

Table 1: Jaccard-based, Pair-wise similarity measure for predicates concerning John Wayne on the imdb dataset.

actedIn	directed	0.100
actedIn	hasWebsite	0.101
bornOnDate	actedIn	0.147
bornOnDate	directed	0.055
bornOnDate	hasChild	0.026
bornOnDate	hasWebsite	0.095
bornOnDate	hasWonPrize	0.086
bornOnDate	isMarriedTo	0.133
directed	hasWebsite	0.025
hasChild	actedIn	0.008
hasChild	directed	0.005
hasChild	hasWebsite	0.013
hasChild	hasWonPrize	0.063
hasWonPrize	actedIn	0.028
hasWonPrize	directed	0.027
hasWonPrize	hasWebsite	0.027
isMarriedTo	actedIn	0.036
isMarriedTo	directed	0.016
isMarriedTo	hasChild	0.083
isMarriedTo	hasWebsite	0.033
isMarriedTo	hasWonPrize	0.135

and $b =$"directed" (i.e. entities that are both actors and directors) and that there are 1000 entities in the dataset that are incident with "actedIn" or "directed" predicate (actors or directors). The similarity measure in such case would have value of: $sim(a,b) = 0.2$.

Table 1 presents the values of similarity measure computed in this way on the imdb dataset for all pairs of predicates incident with the entity *JohnWayne*. For this example, the optimal layout computed by maximising the *vim* measure and using the *sim* measure as described is presented on Figure 2. This example is given only as an illustration of the discussed concepts and definitely leaves a lot of room for improvements. For example, we observed that the proposed simple *sim* similarity measure gives un-intuitive values for some pairs (e.g. "isMarriedTo", "hasWonPrize"), however is generally very promising, considering its simplicity. Also the proposed variant of the *vim* visual integrity measure should be treated only as a basis for further improvements.

3 CONCLUSIONS

We introduced a novel notion of *integrity* in the context of diversity-aware information selection and visualisation tasks and illustrated it on an example of semantic entity summarisation problem. As diversity-awareness has proved to be an important approach in many applications we argue that integrity-awareness is a necessary next step to improve this approaches.

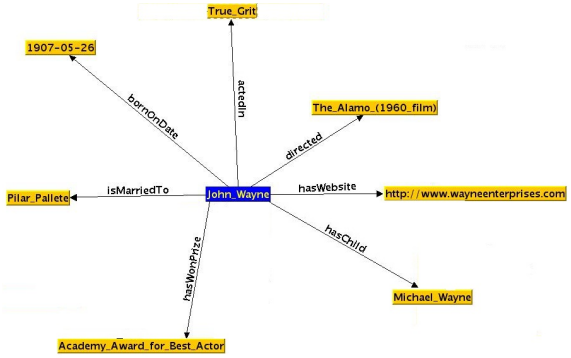

Figure 2: Optimal layout of graphical entity summary of John Wayne computed on imdb knowledge graph with k=7. Integrity measure for this layout L: $VIM(L) = 0.62$.

ACKNOWLEDGEMENTS

The work is supported by Polish National Science Centre 2012/07/B/ST6/01239 "DISQUSS" grant. Thanks are due to M.Andruchów for computing the example in Table 1 and Figure 2.

REFERENCES

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Chen, H. and Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 429–436, New York, NY, USA. ACM.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA. ACM.

Gollapudi, S. and Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, USA. ACM.

Sydow, M., Pikula, M., and Schenkel, R. (2013). The notion of diversity in graphical entity summarisation on

semantic knowledge graphs. *Journal of Intelligent Information Systems*, 41:109–149.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S. A. (2008). Efficient computation of diverse query results. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 228–236, Washington, DC, USA. IEEE Computer Society.

Xu, D., Cheng, G., and Qu, Y. (2014). Preferences in wikipedia abstracts: Empirical findings and implications for automatic entity summarization. *Information Processing and Management*, 50(2):284 – 296.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS