Towards the Foundations of Diversity-Aware Node Summarisation on Knowledge Graphs

Marcin Sydow

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
Polish-Japanese Institute of Information Technology, Warsaw, Poland,
msyd@poljap.edu.pl

Abstract. This paper aims at initiating a discussion of the foundations of the notion of diversity in a novel problem of computing graphical node summarisations in knowledge multi-graphs (equivalently viewed as RDF-graphs). As it reports an ongoing work, it proposes a general framework of basic concepts and adaptations of two diversity-aware evaluation measures previously studied in the context of information retrieval to the studied problem and briefly discusses them.

Keywords: diversity, node summarisation, evaluation measures, axioms

1 Introduction

Consider a large knowledge base in the form of a directed multi-graph where nodes represent entities from some domain and directed arcs represent some binary relations between the entities. For example, in the movie domain, a directed arc labeled as "acted in" could point from the node labeled as "Woody Allen" to the node labeled as "Zelig" (the title of a movie). Assume that arcs have associated numerical weights that represent some notion of "strength" of the particular relation instance.

Now, imagine the task of local, graphical *summarisation* of a particular node q in such a graph G, i.e. the task of extracting a connected subgraph of G surrounding q, that conveys as complete information about q as possible but has a very limited size.

In this paper, we propose a view on this problem that is based on an analogy with information retrieval as follows. The node q can be viewed as a "query" and the elements of the graph G (e.g. nodes and arcs) as pieces of information ("quasi-documents") to be included in the summary based on their "relevance" to the summarised node.

In this context, the simplest approach to construct a summary seems to greedily select the elements of the surrounding graph in the order of their relevance until the size limit is reached. Such approach has a clear analogy with the PRP principle in IR [11].

Actually, [13] has recently presented a greedy algorithm for computing arc-number-limited entity summarisation on RDF-graphs that works in this way, by selecting edges based on their weighted distance from the summarised node. However, the experimental results in [13] revealed that this approach has the problem of high risk of *redundancy of the information* in the summary (such as the dominance of a single relation name), a problem that is inherent to any greedy, PRP-based approach of this kind known in IR. Due to this, [12] proposed "DIVERSUM", a *diversity-aware* variant of the problem

and appropriate novel algorithm that is diversity-aware, in a simple and intuitive way, by explicitly avoiding edge-label repetition when selecting the edges based on their relevance (proximity) and importance (multiplicity). Furthermore, the recent user evaluation study [14] demonstrated that diversity-awareness introduced in the novel algorithm has been clearly appreciated by the users and resulted in higher-quality summaries.

Thus, as explained above, the diversification approach has *a very natural novel application to the problem of node summarisation in knowledge multi-graphs*, an application whose foundations have not been deeply discussed yet, up to the author's best knowledge. Hence, the focus of this paper is to *initiate the discussion of the theoretical foundations of the concept of diversity in node summarisation*. A desired long-term goal of such a discussion is to propose diversity-aware evaluation measures that would make it possible to design diversity-aware node-summarisation algorithms in a more principled way than was done previously (e.g. in [13, 12]).

Contributions: building on [13, 12, 14], we propose an IR analogy to graphical node summarisation, define the general framework of the basic concepts, consider three diversity axioms and propose two (implicit and explicit) diversity-aware evaluation measures for the problem adapted from measures known in IR and briefly discuss them.

Related Work: [9] proposes a random-walk-based summarisation of an information network (not a single node) that is diversity-aware. [10] studies summarisation of tree-structured XML documents within a constrained budget.

Text summarisation is a mature field, see the [8] survey, for example. The issue of diversity has recently gained much interest in IR community. The fact that the relevance of each retrieved document should be evaluated *dependently* on the other retrieved documents was noticed quite early [6]. An early practical diversity-aware re-ranking algorithm, MMR, utilising so called *implicit* (similarity-based) approach to diversification was proposed in [2] and then became a basis for many followers. The problem of diversity naturally appears in the context of *ambiguous* search queries. [3] studied a related problem of providing at least one document relevant to an ambiguous user query. [5] is an example of an *explicit* approach, that directly models various *aspects* of an *under-specified* query, by means of *information nuggets* and proposes a diversity-aware evaluation measure α-nDCG (later combined with other measures in [4]). [1] proposed a category-based model, "intent-aware" evaluation measure and a greedy algorithm approximately optimising it.

2 Generic Specification of the Problem

The underlying knowledge base is a directed multi-graph G with *unique* labels on nodes (representing entities) and (non-uniquely) labelled arcs (representing binary relations) and rational, non-negative weights on arcs reflecting their "strength".

INPUT: a node q to be summarised and $k \in N$, the limit on the summary's S size defined by the function $l(S) \in N$. We consider the size constraint for practical reasons, due to the limited user comprehension capacity or/and the limited display space (especially important in the context of potential applications on small, mobile devices).

OUTPUT: a (weakly) connected subgraph S of G containing q that *satisfies the size constraint:* $l(S) \leq k$ and *maximises* the properly defined *evaluation measure* $f(S)$. In this paper, we discuss the desired properties and propose specific choices for the evaluation measure f.

Considering the definition of the size function l of the summary S, the potential most natural choices are: number of arcs in S (l_a), number of nodes in S (l_n), sum of the two numbers (l_s). Another, more complex choice for the limit function could be the total length l_t of the textual labels of nodes and arcs in the summary.

We introduce a helper operational notion of *information piece*, inspired by the notion of information nuggets in [4], that represents the unit of information contained in S and view S via the notion of D_S i.e. the *collection of its information pieces*. There is some choice on what to consider a single information piece of S: only nodes, only arcs, unique arc labels (relation names), nodes and arcs, nodes and relation names, etc.

Another important concept related to our problem that should be specified is the notion of "relevance" of information pieces in D_S to q. The simplest approach is to define a function $w : D_S \to [0, 1]$ that represents relevance to q, and the total relevance of S is aggregated over the elements of D_S. The relevance function $w(d)$ can have two components: dynamic (i.e. query-dependent) and static (query-independent). Considering the dynamic component, it should take into account: 1) proximity of information piece d to q in terms of the structure of the underlying graph G (e.g. minimum weighted path length from q to d); and 2) similarity between d and q (e.g. based on textual similarity of the labels or other, more sophisticated notions of similarity, based on some ontology, for example). Considering the static component of w, it can be based on some global properties of the graph G such as centrality or prestige measures, etc. known in the field of social network analysis. Due to space limitations we leave a detailed discussion on how to compute $w(d)$ in multi-graphs for future extension of this work.

Notice our assumption of the connectedness of the resulting summary S that literally means that only those information pieces that are connected to q by a path in G could be considered potentially relevant to q.

Specification of the size function $l(S)$, relevance and similarity functions, and decision of what is to be considered a single information piece, seems to be necessary to start a general discussion of evaluation measures for the node summarisation problem.

3 "Axioms" of Diversity-Aware Evaluation Measures

We propose adaptations of 3 "axioms" out of 8 discussed in [7] (for the context of document retrieval) that could be considered in the context of graphical node summarisation.
1) *monotonicity*: $f(S) \leq f(S')$ for any summaries S, S' such that $D_S \subseteq D_{S'}$ (i.e. adding a piece of information to a summary cannot make it worse);
2) *consistency*: the optimal summary S (according to f) does not change if we make its information pieces D_S more relevant to q and less similar to each other and/or other pieces (outside of D_S) less relevant to q and more similar to each other (this definition is valid only if the relevance $w : D_S \to [0, 1]$ and similarity $s : D_S^2 \to [0, 1]$ functions are defined, we will call it *relevance-consistency* if only w is defined).
3) *stability*: $S \subseteq S'$ for any optimal (according to f) summaries S, S' such that S' has larger size than S (i.e. $l(S) \leq l(S')$). Actually, this property is quite strong and it is not clear if it is really desired for a good evaluation measure. It is possible to imagine reasonable examples when an optimum summary containing 2 information pieces is *not* contained in an optimum summary containing 3 information pieces, etc. On the other hand, this property makes it possible that a greedy algorithm that iteratively selects information pieces to add into the summary can find a global optimum.

4 Diversity-Aware Evaluation Measures

We propose two evaluation measures for the node summarisation problem.

DIVERSUM-Based Evaluation Measure: an "implicit" measure that aims at generalising the approach taken in [12] in the form of a convex combination that directly balances the total relevance and maximum allowed redundancy among information pieces:

$$f(S) = \lambda \sum_{d \in D_S} w(d) - (1 - \lambda)|D_S| \max_{d,d' \in D_S, d \neq d'} s(d, d')$$

$\lambda \in [0, 1]$ is a parameter that controls the balance, $w(d)$ is relevance of d to q and s is a pairwise similarity function among information pieces of D_S. $|D_S|$ coefficient stands for balancing the number of terms in the sum. Alternatively, other aggregation functions can be used instead of sum or max. For example, instead of max, it seems reasonable to use sum (with another normalising coefficient: $\frac{2}{|D_S|-1}$ to account for the number of unordered pairs compared to single elements in D_S, in this case we assume $|D_S| > 1$). Using max instead of the first sum is not a good idea since it would not prevent against the "topic drift", i.e. adding irrelevant pieces to the summary. We conjecture that both variants of the measure satisfy consistency but are not monotonic or stable.

Category-Aware Evaluation Measure: (adapted from [1] to our context) it explicitly focuses on ambiguity of the user information need by modeling the distribution of *categories* of information that a user wishing to summarise a node q may be interested in. Let C denote the set of possible categories (or interpretations of the query). The distribution of query interpretations over categories is modeled by $P(c|q)$ (with $\sum_{c \in C} P(c|q) = 1$). Similarly, the relevance function $w(d|c)$ is category-aware. The measure can be viewed as the expected (over all possible interpretations c) value of the chance of satisfying the user with *at least one* information piece of the summary that is relevant to q in the context of the *actual* interpretation of their unknown interest:

$$f(S) = \sum_{c \in C} P(c|q)(1 - \Pi_{d \in D_S}(1 - w(d|c)))$$

For computational tractability, the measure implicitly assumes independent relevance $w(d|c)$ of information pieces *conditioned* on the actual category (due to product) but does not assume independent relevance that would obviously be counter-diversity-aware. We conjecture that the measure is monotonic and relevance-consistent but not stable. Now, we briefly present some novel ideas on how $P(c|q)$ or $w(d|c)$ can be computed for graphs. $P(c|q)$ can be pre-computed once for each node as a soft membership measure obtained from applying any soft clustering method to the nodes of the knowledge base G that takes proximity and textual labels into account. Our preliminary idea for computing $w(d|c)$ is to use probability of getting to d with a k-limited random walk starting at q where the transition probability is skewed towards arcs and nodes that are more relevant to the particular category c. Due to space limitations, more detailed discussion is left for the extended version of this paper. There exists an efficient greedy approximation algorithm optimising this measure (what is claimed to be NP-hard [1]).

Future Work: 1) Deeper discussion and analysis of proposed measures and their variants; 2) practical ways of computing all their ingredients and effective algorithms. Since some of the above measures lead to non-trivial combinatorial search problems, one can consider brute force (for small size of the summary) or some sub-optimal optimisation heuristics (such as simulated annealing, for example); 3) Extensive experimental evaluation of the proposed methods on real data; 4) Further discussion of diversity "axioms".

Acknowledgements. The author is supported by N N516 481940 and N N516 443038 grants of Polish Ministry of Science and Higher Education.

References

1. Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.
2. Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.
3. Harr Chen and David R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 429–436. ACM, 2006.
4. Charles Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rger, Milad Shokouhi, Dawei Song, and Emine Yilmaz, editors, *Advances in Information Retrieval Theory, Proceedings of ICTIR 2009*, volume 5766 of *Lecture Notes in Computer Science*, pages 188–199. Springer Berlin / Heidelberg, 2009.
5. Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.
6. W. Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73–78, 1964.
7. Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 381–390, New York, NY, USA, 2009. ACM.
8. Karen Sparck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449 – 1481, 2007. Text Summarization.
9. Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proc. of the 16th ACM KDD Conference*, KDD '10, pages 1009–1018, New York, NY, USA, 2010. ACM.
10. Maya Ramanath, Kondreddi Sarath Kumar, and Georgiana Ifrim. Generating concise and readable summaries of xml documents. *CoRR*, abs/0910.2405, 2009.
11. S. Robertson. The probability ranking principle. *J. of Documentation*, 33(4):294–304, 1977.
12. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. DIVERSUM: Towards diversified summarisation of entities in knowledge graphs. In *Proceedings of Data Engineering Workshops (ICDEW) at IEEE 26th ICDE Conference*, pages 221–226. IEEE, 2010.
13. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. Entity summarization with limited edge budget on undirected and directed knowledge graphs. *Investigationes Linguisticae*, 21:76–89, 2010. `http://www.staff.amu.edu.pl/~inveling/index.php?direct=227`.
14. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. To diversify or not to diversify entity summaries on RDF knowledge graphs? In *(to appear in) The Proceedings of the ISMIS 2011 Conference*, Lecture Notes in Artificial Intelligence. Springer Verlag, 2011.

Appendix
(Drafts of Selected Proofs)

This appendix contains drafts of selected proofs of the statements concerning the axiomatic properties of the discussed evaluation measures proposed in section 4. The refinements of the drafts and other proofs are left for the extended version of this paper.

I) DIVERSUM-Based Evaluation Measure: We show drafts of the proofs that the measure satisfies consistency but does not satisfy monotonicity.

We consider both the variants proposed in the paper:

a) $f(S) = \lambda \sum_{d \in D_S} w(d) - (1-\lambda)|D_S| \max_{d,d' \in D_S, d \neq d'} s(d,d')$

b) $f(S) = \lambda \sum_{d \in D_S} w(d) - \frac{2(1-\lambda)}{|D_S|-1} \sum_{d,d' \in D_S, d \neq d'} s(d,d')$

Let S denote the optimal summary subgraph according to $f(\cdot)$

Consistency: if the information pieces in D_S become more relevant to q (or stay the same) and less similar to each other (or stay the same) and the opposite change concerns the information pieces (or pairs) not inside D_S (or they do not change) the value of the first term in the formula for $f(S)$ increases and its second term decreases so that their difference increases (or stays the same) for any value of λ. Moreover, there is no such increase in the value of $f(S')$ for any other subgraph $S' \neq S$. Thus, S stays an optimal choice (notice that the definition of consistency implicitly assumes that there is a unique optimal subgraph S).

Monotonicity: for showing that the measure is not monotonic, denote by S some node summary and consider an information piece $\delta \notin D_S$ that has the following property.

Ad a) $\exists_{\delta' \in D_S}$ s.t. $s(\delta,\delta') > \max_{d,d' \in D_S, d \neq d'} s(d,d')$ and $s(\delta,\delta') > \frac{\lambda}{(1-\lambda)|D_S|} w(\delta)$.

It holds that $f(S \cup \{\delta\}) < f(S)$ for such δ, which contradicts monotonicity.

Ad b) to contradict the monotonicity it is necessary that (below, $\delta = |D_S|$):

$(\delta - 1) \sum_{d,d' \in D_S \cup \{\delta\}, d \neq d'} s(d,d') - \delta \sum_{d,d' \in D_S, d \neq d'} s(d,d') > \frac{\lambda}{1-\lambda} \frac{\delta(\delta-1)}{2} w(\delta)$

which is possible if the δ novel terms present in the first sum amount for a total similarity that is high enough to balance the $\delta/(\delta-1)$ coefficient ratio between the first and the second sum and the value on the right side of the inequality.

II) Category-Aware Evaluation Measure:

$$f(S) = \sum_{c \in C} P(c|q)(1 - \Pi_{d \in D_S}(1 - w(d|c)))$$

Relevance-consistency: The measure is obviously relevance-consistent, since it is an increasing function of all the values of the conditional relevances $w(d|c)$ of information pieces $d \in D_S$.

Monotonicity: The measure is monotonic since extending the summary with a new information piece $\delta \notin D_S$ introduces one more multiplicative factor $0 \leq 1 - w(\delta|c) \leq 1$ to the product of each term of the sum in the formula that makes that the total value of $f(S)$ is not lower than before the extension.

The proofs of stability are left for the extended version of this paper.