J Intell Inf Syst (2013) 41:109–149
DOI 10.1007/s10844-013-0239-6

# The notion of diversity in graphical entity summarisation on semantic knowledge graphs

**Marcin Sydow · Mariusz Pikuła · Ralf Schenkel**

Received: 3 July 2012 / Revised: 12 February 2013 / Accepted: 13 February 2013 /
Published online: 12 March 2013

**Abstract** Given an entity represented by a single node $q$ in semantic knowledge graph $D$, the Graphical Entity Summarisation problem (GES) consists in selecting out of $D$ a very small surrounding graph $S$ that constitutes a generic summary of the information concerning the entity $q$ with given limit on size of $S$. This article concerns the role of *diversity* in this quite novel problem. It gives an overview of the diversity concept in information retrieval, and proposes how to adapt it to GES. A measure of diversity for GES, called ALC, is defined and two algorithms presented, baseline, diversity-oblivious PRECIS and diversity-aware DIVERSUM. A reported experiment shows that DIVERSUM actually achieves higher values of the ALC diversity measure than PRECIS. Next, an objective evaluation experiment demonstrates that diversity-aware algorithm is superior to the diversity-oblivious one in terms of fact selection. More precisely, DIVERSUM clearly achieves higher recall than PRECIS on ground truth reference entity summaries extracted from Wikipedia. We also report another intrinsic experiment, in which the output of diversity-aware algorithm is significantly preferred by human expert evaluators. Importantly, the user feedback clearly indicates that the notion of diversity is the key reason for the

M. Sydow (✉) · M. Pikuła
Web Mining Lab, Polish-Japanese Institute of Information Technology, Warsaw, Poland
e-mail: msyd@poljap.edu.pl

M. Pikuła
e-mail: mariusz.pikula@poljap.edu.pl

M. Sydow
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

R. Schenkel
Saarland University and MPI for Informatics, Saarbrücken, Germany
e-mail: schenkel@mmci.uni-saarland.de

preference. In addition, the experiment is repeated twice on an anonymous sample of broad population of Internet users by means of a crowd-sourcing platform, that further confirms the results mentioned above.

**Keywords** Diversity · Semantic knowledge graph · Graphical entity summarisation · Summarisation · Diversity measure · Algorithm · Experiment · Crowdsourcing

## 1 Introduction

Semantic knowledge graphs are gaining importance in research and applications especially in the presence of the systems that automatically harvest semantic knowledge from the open-domain sources such as WWW, for example.

In such graphs, the nodes represent entities (e.g. in the movie domain: actors, directors, or movies) and the directed labeled arcs represent relations between entities (e.g. "acted in", "directed", "has won prize"). Formally, they are multi-graphs, since multiple arcs incident with a single pair of nodes are possible as there are possible instances of multiple relations between two entities (e.g. a person directed and acted in a film).

Each arc of such a graph together with its endpoints can be interchangeably called an "s-p-o triple" for it consists of two nodes representing *subject* and *object* connected by an arc representing a *predicate* that specifies the relationship between the entities represented by the nodes.

More importantly, such an arc naturally represents a simple atomic *fact* that has some semantic meaning in the domain represented by the graph.

Figure 1 presents a small semantic knowledge graph being a fragment of a dataset extracted automatically from the IMDB portal,[1] concerning the movie domain.

Arcs in the knowledge graph can have numerical weights that reflect some additional information concerning the represented fact, for example importance (that may be represented by "witness count" – reflecting the number of documents in a base corpus that contained the fact represented by the arc), credibility, novelty, etc.

### 1.1 Motivation

Although semantic knowledge graphs represent partially structured[2] type of data, and there exist some structured query languages for querying them (e.g. SPARQL)[3] there is an increasing interest in *unstructured querying* (e.g. keyword querying) over such data model. Unstructured querying is easier to apply for inexperienced users, does not demand any prior familiarity with the structure of the underlying semantic knowledge graph and is generally more natural for humans to use. Thus, it bridges the world of structured languages such as SQL for standard databases with the unstructured world of IR (information retrieval) with its more relaxed querying based, for example, on textual keywords.

[1] www.imdb.org

[2] Sometimes called *semi-structured* as opposed to "fully" structured relational databases, for example.

[3] http://www.w3.org/TR/rdf-sparql-query/

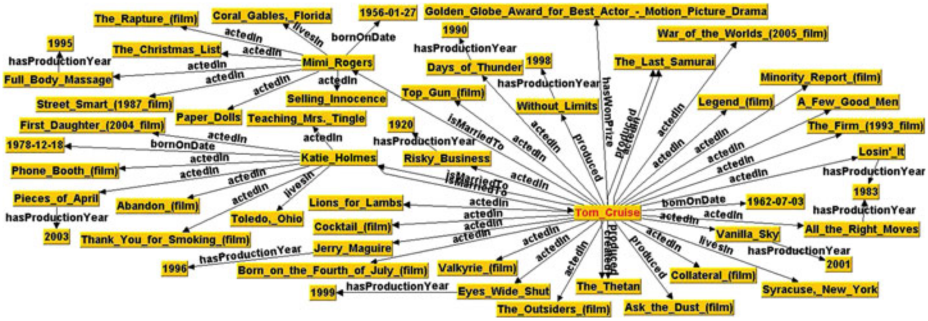


**Fig. 1** An example of a fragment of a semantic knowledge graph concerning movie domain, extracted from the IMDB database and visualised by our tool. It is a radial neighbourhood of the node representing an actor "Tom Cruise" in the distance of $r = 3$ hops around the entity

The most basic unstructured query that can be imagined in the context of semantic knowledge graphs could be like *"tell me a few most important facts about given entity"*.

The connected component of the semantic graph containing an entity contains information that potentially concerns it. This information can be viewed as a composition of facts (i.e. s-p-o triples). Due to the very nature of semantic knowledge graphs, the most relevant facts about an entity in such a graph are in the topologically closest neighbourhood to the node representing the entity.

However, as one can easily notice on Fig. 1, even for a very small radius around an entity, there are definitely too many facts to be useful as a small, generic summary for a human.

Henceforth, it would be very valuable to have a tool to efficiently select only a very limited fragment of such a graph around the node representing given entity as its *summary* that can quickly provide some basic facts about the entity. Such a problem called GES (for: *Graphical Entity Summarisation*) is the main topic of this publication. It was less formally introduced and preliminarily studied in a series of our quite recent publications (Sydow et al. 2010a, b, 2011) whose this article is a substantial extension of.

In the GES problem on semantic knowledge graphs, a user specifies an *entity* in order to let the system to generate a *summarisation of facts concerning this entity*,[4] for instance, an actor "Tom Cruise" that is represented by a node $q$ in the knowledge graph $D$, and expects the system to return its *graphical summarisation*. By graphical summary of an entity we mean a small subgraph $S$ of knowledge base $D$ that contains a limited *selection* of facts that directly or indirectly concern the entity represented by $q$. It may be viewed as a generic summarisation of the information concerning this entity. More formal specification of the GES problem and a wider discussion of the concept of summarisation will be given later in this article.

Naturally, the GES problem can be viewed as a way of *unstructured querying* the knowledge graph, where the label of the node-entity to be summarised can be viewed as a simple, unstructured query $q$ that is sent by the user in order to obtain some

[4] Henceforth, we will call it shortly as "entity summarisation".

*result*. Using this view it is possible to adapt some concepts from IR (information retrieval).

But, even more naturally, it can be interpreted as a novel incarnation of the *text summarisation* problem, that have been heavily studied in the context of summarising *textual documents* before.

Part of contribution of this work, taking apart the novel concept of GES itself, is to synthesise some selected concepts from both disciplines mentioned above and adapt them to the new domain of semantic knowledge graphs.

To illustrate one of potential practical applications of the GES problem, we present the following example scenario. Assume that a user heard on a street a fragment of talk with a single reference to a name of a person, assume that it was "Tom Cruise", for instance. The user has an impression that the person is famous but is not sure who this person is and immediately wants to check it. Assume the user has a mobile phone with installed application that has an access to a semantic knowledge graph concerning the movie domain. The user quickly inputs the name of the entity to the system (by voice, for example) and the result, in the form of GES, is immediately presented on the small display of the device. The user at a glance understands that the person is a famous actor and, in addition, learns a few basic facts about him.

In the scenario above, the important elements are:

- limited possibilities of operating the device (only passing the input phrase by voice, for example) as the user walks the street, for instance
- limited size of the display of the device
- limited possibilities of processing the output: only a glance on the result must be enough to make the result useful, i.e. no detailed reading is possible. In addition, only some very simple manual operations such as touching the screen to zoom or move some elements are possible
- limited amount of time that will be accepted by the user while waiting for the response

We argue that GES is the right solution for the above example scenario, and the existing models, such as search engines do not provide such functionality yet (in particular more reading is necessary to learn basic facts about the searched entity in a search engine). We believe that GES would have many more interesting applications. For example, it could be partially used a submodule in systems that help enriching results presented to the users of search engines, like Google Knowledge Graph, for instance.

In this article all the examples concern actors and movie domain. However, in general, the summary does not necessarily have to concern a human, but any other type of entity (like city, movie, etc.). Also the domain may be different than movies.

The essence of any summary is its compact form. In addition, in the exemplary scenario above we assume that the display to present the result is small.

Because of this, we model the limited comprehension of the user or limited capabilities of the displaying device by introducing the notion of limited *presentation budget* represented by an additional parameter: $k \in N_+$ which is defined in this article as the upper bound on the number of triples (arcs) allowed to be presented in the returned summary.

Finally, and most importantly, we focus in this article on the problem of *diversification* of such summary, which is of importance especially in the context of limited presentation budget.

The aspect of diversification is very important since the knowledge graph is usually created as the integration of non-homogeneous information extracted from various sources (including the Web) so that it can represent many different *facets*, *types* or *categories* of knowledge even if these various categories of information concern the same entity. A particular user might have some specific *interest profile* i.e. be especially interested in seeing some specific categories or facets of information concerning the summarised entity without the necessity to explicitly inform the system about it. Reasons for this would range from preserving privacy/anonymity to limited possibilities of operating the interface (as in the scenario above).

In this article we take the simplistic assumption that types, categories or aspects of information in semantic knowledge graph are fully represented by arc labels. In particular, we model the user interest profile by the set of arc labels she is interested in.

For example, the user may be more interested in the private aspects of the life of an actor (such as when the actor was born, where he lives or who is his wife, for example) rather than his professional activity. Thus, in such a model, the value of a piece of information (fact) to such a user would depend on the arc label i.e. arc labels “bornIn”, “livesIn”, “marriedTo”, would be of interest, unlike “actedIn”, “directed”, etc.

Let’s further illustrate this issue on the same example. In the knowledge base extracted from IMDB database, concerning movies, that we use for experimentation, the node labelled “Tom Cruise” is adjacent to over 30 different arcs[5] (Fig. 1, again).

About 20 of them are labelled “actedIn”, the second dominating group consists of 4 arcs labelled “produced”. Among the remaining few arcs, only 4 concern the private life of the actor (labelled “bornOnDate”, “isMarriedTo” – two arcs, as the actor married more than once – and “livesIn”).

Thus, if the limit $k$ on arcs to be presented is low (e.g. $k = 10$) a diversity-unaware system is very likely to present only the arcs concerning the professional aspects of the entity (26 out of 30 adjacent arcs) leaving the user interested in private aspects dissatisfied with the summary.

Deeper discussion of the diversification concept is given in the Section 3.

We believe that incorporating diversity-awareness into the GES problem would help to improve the results which is one of the main topics of the subsequent sections.

### 1.2 Main contributions

The main contributions of this article are the following:

1. The concept of GES (Graphical Entity Summarisation) and its discussion
2. Formally introducing the notion of diversity into the GES problem by means of the ALC (Arc Label Coverage) measure for the GES problem
3. Experimental comparison of the level of diversity-awareness of two algorithms for the GES problem: PRECIS and DIVERSUM on real data

[5] After reducing the number of relations as explained in Section 5.2.

4. A series of experimental results that clearly indicate superiority of the diversity-aware algorithm over the diversity-oblivious one, in particular, in terms of:
   - better fact selection, as compared with ground truth reference entity summaries from Wikipedia
   - significantly better user assessment of the outputs viewed as generic entity summaries
5. Analysis of collected user feedback that indicates that diversity-awareness is a highly demanded property for the GES problem

The article is a substantial extension of a conference paper (Sydow et al. 2011) that presents initial assessment experiment on the discussed issue. It also builds on our earlier workshop and conference papers (Sydow et al. 2010a, b). In those publications, the graphical entity summarisation problem was originally mentioned in a less formal way and two algorithms (PRECIS and DIVERSUM) for this problem were introduced, together with preliminary experimental results.

### 1.3 Detailed contents of the article

The structure of the article is as follows:

1. (Section 1) discussion of the novel notion of graphical entity summarisation (GES) and example of its possible application
2. (Section 2) the problem of summarisation and recent related work (Section 2.1)
3. (Section 3) the concept of diversity including concise survey on diversification of results in information retrieval (Section 3.1) and ideas on how it can be adapted to graphical entity summarisation (Section 3.2)
4. (Section 4) algorithms for the GES problem, including more formal specification of the GES problem (Section 4.1) , description of two algorithms for GES: PRECIS (diversity-oblivious) (Section 4.2) and DIVERSUM (diversity-aware) (Section 4.3) with an illustration on a toy example (Section 4.4)
5. (Section 5) description of experimental setup including implementation of a prototype experimental platform (Section 5.1), real dataset concerning movie domain (Section 5.2) and the test data sample concerning famous actors (Section 5.4)
6. (Section 6) diversity-awareness experiment confirming higher degree of diversity in DIVERSUM than in PRECIS in terms of ALC – an objective diversity measure defined in Section 6.1
7. (Section 7) intrinsic evaluation experiment based on reference ground truth summaries, indicating superiority of diversity-aware DIVERSUM over diversity-oblivious PRECIS in terms of fact selection
8. (Section 8) expert-based assessment experiment that indicates preference of DIVERSUM over PRECIS including its detailed analysis
9. (Section 8.3) analysis of feedback of the experts that explicitly emphasise the value of diversity-awareness of the outputs

10. (Section 9) additional two crowd-sourcing-based experiments that additionally support the experimental results presented in this article, its analysis and description of the crowd-sourcing approach
11. (Section 10) conclusions and ideas on future continuation work

## 2 Summarisation

Automatic summarisation of textual documents has been intensively studied and is well represented in research literature. We report related work in the Section 2.1.

Text summarisation, after Mani (1999), is the process of *"taking an information source, extracting content from it, and presenting the most important content to the user [...]. Summaries can be user-focused [...] tailored to the requirements of a particular user or group of users, or else, they can be generic, i.e., aimed at a particular – usually broad – readership community."*.

The GES problem studied in this paper can be viewed as a *generic* type of summarisation i.e. aimed at a broad group of users with various information needs for whom the particular task is not known. Of course, GES is not the same problem as text summarisation, but the analogies are obvious and we build the concept of GES on those analogies.

Mani (1999) further distinguishes between *extractive* and *abstractive* summaries. The former one is more basic and is achieved by simply *selecting* (extracting) the elements, typically sentences, from the summarised text to present them as the summary. The latter one, more sophisticated, is capable of synthesising new elements out of the summarised text.

The GES problem naturally corresponds to the extractive generic summarisation as it selects s-p-o triples from the underlying knowledge graph to compose the resulting graphical summary of the entity.

An important issue in text summarisation is how the following three tasks are solved:

- content selection (how to select the elements to compose the summary?)
- information ordering (how to order the selected elements?)
- sentence realisation (how to present the summary to the user?)

Analogous issues concern the GES problem. In this paper we focus mainly on the first issue - how to *select* the triples to compose the summary. Since in what we propose, the resulting summary is in the form of a graph, the two other issues are very interesting and non-trivial.

However, the issue on how to organise and layout the resulting summary graph is out of the scope of this article and deserves a separate publication. Concerning the layout, in the prototype software that we use for experiments, we apply some novel algorithms for automatic layout computation for the GES problem. In particular our novel AGNES visualisation algorithm is reported in Sobczak et al. (2012).

Finally, concerning the evaluation of summarisation, one can consider *intrinsic*, i.e. task-independent, or *extrinsic*, i.e. task-dependent methods. Since we consider the GES problem as a general, *task-independent* summarisation problem rather than a tool for solving any specific user task, we will apply the intrinsic experimental evaluation.

More precisely, we will use for the evaluation an adaptation of the well known ROUGE method, a recall-based approach to evaluate textual summaries (Lin and Hovy 2003). The details of the evaluation methods are given in the experimental sections.

### 2.1 Related work on summarisation

Up to the best of the authors' knowledge the notion of graphical entity summarisation has not been studied in research literature before our work.

The problem studied in this article is cross-disciplinary since it concerns summarisation (that will be discussed in this section) with a special focus on graphs and diversity (that will be discussed in Section 3).

Summarisation has been intensively studied for text documents (Wan and Xiao 2010), and many techniques for extractive and abstractive Summarisation have been proposed; extensive surveys are given in Gupta and Lehal (2010); Hovy (2005); Spärck Jones (2007). The field includes diverse fields such as Summarisation of conversations (Carenini et al. 2011), scientific papers (Abu-Jbara and Radev 2011) , and news articles (Barzilay and McKeown 2005), cross-language Summarisation (Wan et al. 2010), summarising the differences of documents (Wan et al. 2011), multiple documents (Wan 2009) and summarising for non-native speakers (Wan et al. 2010). The problem of increasing diversity of document summaries has been considered in Li et al. (2009) with a focus on reducing redundancy in sentences of a summary. In contrast to this large body of work, our method produces a summary of an entity, not a text.

Zhang et al. (2010) (as a recent example for a large set of similar papers) consider the problem of creating concise summaries for very large graphs such as social networks or citation graphs. Related methods have been proposed for summarising ontologies (Zhang et al. 2007; Li et al. 2010; Cheng et al. 2011), where an important application is generating snippets for semantic search engines (Penin et al. 2008). In contrast to this, our work aims at summarising information around a single node in a graph. Ramanath and Kumar (2009) and Ramanath et al. (2009) propose methods for summarising tree-structured XML documents within a constrained budget.

The problem of entity summarisation in the form of a graph with limited number of edges has only recently been studied. The problem was originally proposed in Sydow et al. (2010b), together with an efficient algorithm, called PRECIS. RELIN (Cheng et al. 2011) tackles a similar problem in the sense that it computes entity summaries with a limited size, using a random walk on a graph of features characterising the entity. However, its main purpose is a quick identification of the entity, so it focuses on selecting distinctive information, not important information in general. Furthermore, it does not consider diversity of the summary, and it does not consider properties of related entities.

Waitelonis and Sack (2012) consider entity Summarisation in the context of exploratory search. They discuss a number of heuristics for ranking properties of DBPedia entities, with a focus on identifying properties that are connected to related entities. Thalhammer et al. (2012) first exploit usage information (such as ratings etc.) to identify, for the entity to summarise, the $k$ most similar entities in the collection. The summary then contains properties that are frequently shared with the $k$ neighbours, but not with the remaining entities. Their focus is on finding properties that are important for this group of entities. None of these methods considers the diversity of the resulting summary. Finally, entity Summarisation is also an ingredient of the Google Knowledge Graph[6] that can deliver, for some queries, important facts of an entity instead of the usual list of links.

In a recent work, Thalhammer et al. (2012) propose to assess the quality of an entity summary using a game-based ground truth. The interestingness of a property is determined as its popularity among the game players, computed as the fraction of game players who could answer a question about a fact with that property. In their evaluation, they used 60 common movies from IMDB to compare the system from Thalhammer et al. (2012) with the Google Knowledge Graph, but did not find a significant difference of the two.

There is a large body of work on visualising, exploring, and analysing large graphs, for example (Kairam et al. 2012) for networks, (Wattenberg 2006) for multivariate graphs, (Ham et al. 2009) for social networks, and (Dokulil and Katreniaková 2008) for RDF graphs. None of these works considers presenting only selected edges of a graph node (which would correspond to our entity summarisation problem).

Li et al. (2010) consider the generation of natural language summaries for entities based on template sentences extracted from existing textual summaries of other entities. This results in rather regular summaries for entities of the same type, whereas our work constructs a summary with the most important facts for each entity. Generating natural language summaries is orthogonal to our work, and we could apply some of the techniques discussed in this paper for an improved result presentation (which is beyond the scope of this article).

## 3 Diversity

The concept of *diversity* is of increasing practical importance in various fields such as information retrieval, recommender systems, databases, etc. In this section we focus mainly on the application of this concept in information retrieval, where it has been most intensively studied recently. In Section 3.2 we propose how to adapt it to the task of graphical entity summarisation.

One of the key challenges in information systems that involve returning some results to unknown users is how to satisfy the user's unknown information need that is hidden behind the ambiguous and underspecified query, especially when the personal profile of the user is not available.

For example, in the case of web search engines, an average user submits a very short query that can have many interpretations and aspects. At the same time, a typical user is capable of inspecting only a very small number $k$ of top returned

[6] http://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html

results (in practice *k* is many orders of magnitude smaller than the total number of documents in the system, typically it is not higher than 10 for an average user).

A natural and simple approach to select the *k* of results to be presented to the user (the approach that dominates in the earlier generations of search systems) is to assign some numerical measure of "relevance" to the user query to each potential result kept in the repository of the system and return the top-*k* results according to this measure. This approach is based on the classical PRP assumption (Probability Ranking Principle (Robertson 1977)) and makes it possible to apply simple and fast algorithms to build the result set returned to the user.

It is easy to observe that the results constructed with such an approach can be of poor quality to the user for the following reasons:

- the items in the result set may be very similar to each other, i.e. one may observe *high redundancy* of the result set, which is a problem because of the very limited size of the set.
- the result set can be dominated by items that represent the "most relevant" or "most popular" interpretation or aspect of the information need that might present no value for some less typical users

These problems are due to the fact that each element of the result set is selected separately based only on its relevance to the query with no control over inter-item similarity. We will illustrate this issue with the ambiguous and multi-aspect query "windows". This query has many potential *meanings*, for example the name of a popular operating system, or the plural form of an opening in the wall of a building, etc. Furthermore, even if the actual user's intention of the meaning of the query is known (what is usually not the case in practice), there are various possible *aspects* of each meaning (e.g. "where to download the system", "how to install it", "what are available versions", etc.). In this situation, the application of the classic PRP-like approach to collect the results separately, according to their individual relevance to the query, would likely omit some important interpretations or aspects; often, the results may be dominated by the most popular interpretation/aspect.

This problem was early noticed in information retrieval, e.g. Goffmann (1964) observes that the relevance value of each item returned to the user should depend on the other returned items. A remedy for the problems presented above, that has recently gained an increasing interest, is to introduce some controlled level of *diversity* to the results presented to the user, i.e. to potentially sacrifice a small part of individual relevance of the returned items in order to increase their variety—an approach very different to the classic PRP assumption. In this way, it is possible to

- reduce the redundancy in the result set (to not "waste" the limited number of available slots in the result set), and
- cover more potential interpretations or aspects of the user's unknown information need to minimise the risk of retrieving *no* item related to the actual user information need among the top-k results.

We argue in this article that, by analogy to information retrieval, graphical entity summaries would prove much more useful for the users when the concept of diversity is introduced.

As the experimental results later in this article (Sections 8, 9) show, this thesis can be proven empirically, i.e. users are significantly more satisfied with the entity summarisations that are diversified.

### 3.1 Selected diversity-aware approaches in IR

In this subsection, we present an intentionally restricted sample of diversity-aware approaches that have been successfully used in information retrieval. The ideas presented in those works can serve as a basis for adaptation in the diversity-aware approach to the GES problem. An extensive survey of works concerning application of diversity in IR deserves a separate publication and is out of the scope of this article, due to space limitations.

One of the first practical applications of diversity to improve the quality of the results presented to the user is Carbonell and Goldstein (1998). It proposes a new ranking criterion called *MMR* (for "Maximal Marginal Relevance") that represents an explicit compromise between relevance and diversity among the result set. The measure is a parametrised linear combination of "relevance" and "novelty". That work also reports a simple experimental user evaluation that indicates that the MMR method is preferred by the users, compared with a traditional, relevance-based ranking method, in the context of textual information retrieval and document summarisation tasks.

Under the suggestive title "Less is more" Chen and Karger (2006) define the diversity-aware approach as maximisation of the probability that *at least one* result among the top-$k$ is relevant. Notice that this approach differs from the classic approach of PRP that maximises the number of relevant documents among the top-$k$ results, i.e. maximises the precision. After simplification their model leads to a greedy algorithm that selects the next document in the presented list assuming that all previous (unranked) documents were not relevant.

Clarke et al. (2008) propose a diversity-aware evaluation measure $\alpha - NDCG$ that accounts for high coverage of different aspects of the information need, represented as "information nuggets". The notion of information nuggets was adapted from previous uses in text summarisation and query answering. The proposed measure extends NDCG (normalised discounted cumulative gain) – a diversity-unaware evaluation measure known in information retrieval. The proposed measure has a parameter for modelling human judgement errors.

Another diversity-aware approach, coined "Intent-aware" (IA), is presented in Agrawal et al. (2009). It is stated as the maximisation of the probability of *satisfying an average user* under known distribution of possible interpretations of a given query. More precisely, it concerns re-ranking of the search results in the context of an ambiguous query, assuming the existence of a category hierarchy: the query can belong to one of many categories according to a distribution that is known. Additionally it assumes that the probability of relevance of a document to a query conditioned on its category is known. The problem is formulated as finding the set $S$ of $k$ documents maximising the probability that at least one document will be relevant to an average user. As the paper explains, the problem is NP-hard (by reduction from MAX COVERAGE) but the authors also prove that the objective function is submodular, allowing for a greedy approximation algorithm (called IA-Select in the paper) that has $1/(1-e)$ approximation guarantee. That paper also proposes three "intent aware" extensions of classic IR evaluation measures: NDCG, MRR (mean reciprocal rank), and MAP (mean average precision). The proposed measures are basically expected values of the original measures with regard to the query category distribution. The paper reports an extensive experimental evaluation based on proprietary datasets and human evaluations, in the latter case with help of the crowd-sourcing Amazon's Mechanical Turk platform. The experimental results show that the proposed solution consequently outperforms top three search engines measures with all the proposed intent-aware evaluation metrics.

Gollapudi et al. (2009) use the axiomatic approach to characterise and design diversification systems. They develop a set of axioms that a diversification system is expected to satisfy, and show that no diversification function can satisfy all these axioms simultaneously. Finally, they propose an evaluation methodology to characterise the objectives and the underlying axioms. They conduct a large scale evaluation based on Wikipedia and a product database.

The issue of result diversification in structured data, e.g. in relational databases was studied in Vee et al. (2008).

## 3.2 The concept of diversified GES as an adaptation from IR

This section proposes ideas on how the key concepts in IR (information retrieval) can be adapted to naturally define the basics of the concept of graphical entity summarisation, including *diversity* and hints on how to practically solve the GES problem.

Generally speaking, classic IR deals with the following task. Given an implicit user information need $I$ that is imperfectly represented by a keyword-based query $q$ and a corpus of textual documents $D$ the task is to return the (ordered) set $S \subseteq D$ of documents containing information that satisfies $I$. The following criteria are usually taken into account when computing the result set $S$ in IR:

- *relevance* (how much the results are relevant to the query?): in IR it is based on textual or semantic similarity of the results (documents) to the query
- *importance* (how important, independently on the query, the results are?): in Web IR it can be based on link-analysis of the hyperlink graph (e.g. PageRank)
- *popularity* (how popular are the results?): in Web IR it can be based on user-behaviour recorded in logs (e.g. clicks or viewing times)
- *diversity* (how diverse are the results to satisfy potentially very different information needs behind the query?): it means to cover in the result set as many different aspects of the query as possible and avoid redundancy in the results

Looking from the IR perspective we can view the GES problem in an analogous way, with the following adaptations:

- "query" $q$ corresponds to the node representing the entity to be summarised
- "document corpus" $D$ corresponds to the underlying semantic knowledge graph and consists of facts (triples) represented by arcs
- "result set" $S$ is a subset of selected facts (triples) that constitute a summarisation of the information concerning the entity $q$.

Now we propose to adapt the four criteria listed above to be considered while selecting facts to the summary in the GES problem.

We also propose the following ways of relating these criteria to the domain of semantic knowledge graphs:

- *relevance*: topological proximity of an arc to the query-entity in the semantic knowledge graph. I.e. we assume that the arcs that are less number of hops from the summarised node represent facts that are more "relevant" to this node. In particular, the arcs that are incident to the node are the most "relevant".
- *importance*: based on the arc weights that represent "importance" of facts and are obtained during the knowledge harvesting procedure
- *popularity*: based on the statistical frequency of arc label in the topological neighbourhood of the summarised node in the graph. I.e. more frequent arc labels are "more popular". In this way, if an entity is, for example, an active actor, and not a director, "actedIn" arc label is expected to be "more popular" arc label concerning this entity than "directedIn"
- *diversity*: high coverage of different arc labels in the summary to represent different aspects of information concerning the summarised entity

The set of proposed four criteria (or its subsets) will be used in this article to design concrete algorithms for computing the GES problem in Section 4.

More precisely we will propose two algorithms, one (baseline), diversity-oblivious satisfying only the first two criteria above and the second satisfying all the criteria, notably including the *diversity* criterion.

One of the aims of this article is to experimentally compare the performance of the algorithms and to check whether including the diversity criterion:

- improves the fact selection in the GES problem (experiment in Section 7)
- makes the results more appreciated by the users (experiments in Sections 8, 9)

## 4 Algorithms

In this section we give the algorithmic specification of the GES problem on knowledge graphs and describe two algorithms for this problem, a diversity-oblivious one that does not take into account the diversity criterion (defined in Section 3.2) and a diversity-aware one that takes the criterion into account.

### 4.1 GES problem specification

The problem of graphical entity summarisation can be specified as follows:

**INPUT:**

1. $D$ – an underlying knowledge base (a directed multi-graph) with labels on arcs (representing instances of binary relations between entities) with positive real weights on arcs reflecting "importance" of arcs
2. $q$ – a node of $D$ (entity to be summarised)
3. $k \in N$ – an upper limit on the number of facts (triples) in the summary, also referred to as "presentation budget" in this article.

**OUTPUT:** $S$ – a connected subgraph of $D$ containing $q$ and at most $k$ arcs that together represent a collection of facts being a summary of information concerning the entity in the semantic knowledge graph.

Summaries will be constructed with regard to the set of criteria defined in Section 3.2 reflecting the notions of “relevance”, “importance”, “popularity” and “diversity”.

**Weights on arcs and the notion of “distance”** Notice the assumption that arcs are annotated with real-valued, positive weights that represent “importance” (or “strength”) of the arc. Such weights are treated in this article as an external input. For example, in the datasets referred to later in the article the importance is based on the “witness count” value. It is computed during the knowledge harvesting procedure, that is out of scope of this article, and reflects the number of times the fact represented by the arc was found in the documents of the processed base text corpus.

Intuitively, the criterion of “importance” implies the preference of the arcs with high “witness count” value. We apply a simple technical trick in this article. Namely, we substitute the value of “witness count” with its *inverse* and call it “distance”. Thus, maximising “importance” is equivalent to minimising “distance”. In this way, it is possible to naturally adapt graph algorithms that greedily select the elements in the smallest distance from the starting node to achieve the goal of selecting the most “important” facts. Furthermore, by introducing the notion of distance in this way it is possible to propagate the notion of importance for arcs that are more than one hop away from the summarised node. We naturally define an “aggregated distance” for such arcs as the minimum possible sum of “distances” (i.e. inverses of “witness count”) of arcs forming a path from the summarised node $q$ to this arc, inclusively.

## 4.2 Baseline (diversity-oblivious) algorithm PRECIS

We describe an efficient greedy algorithm for the GES problem, called PRECIS.

The idea of the algorithm follows the analogy with the PRP principle (Robertson 1977) in IR. The facts (arcs) to the summary are collected according to the following two criteria:

- (first) *relevance* (arcs topologically close to the summarised node)
- (second) *importance* (arcs having high “witness count” or, equivalently, having low “distance” to the summarised node, as explained in the previous subsection)

More precisely, it constructs the summary starting from the summarised node “q” by greedily selecting the topologically closest (relevant) edges in the order of low “distance” until $k$ edges are selected. Due to this idea, the algorithm can be viewed as an adaptation of the classic Dijkstra algorithm for computing shortest paths (Cormen et al. 1990) from a single source (summarised entity) to the nodes in the base graph

*D*. The algorithm terminates when up to *k* arcs are collected. This is presented in Algorithm 1.

**Algorithm 1** The PRECIS algorithm for computing entity summarisation

```
PriorityQueue $PQ$
Set $RESULT$
**for all** $a$ in radius $k$ from $q$ **do**
    $a.weight \leftarrow 1/witnessCountOf(a)$
    $a.distance \leftarrow "infinity"$
**end for**
**for all** $a$ adjacent to $q$ **do**
    $a.distance \leftarrow a.weight$
    $PQ.insert(a)$
**end for**
**while** $RESULT.size \leq k$ **and** $(currentArc = PQ.delMin()) \neq null$ **do**
    **for all** $a$ in $currentArc.adjacentArcs$ **do**
        **if not** $RESULT.contains(a)$ **then**
            $a.distance \leftarrow min(a.distance, (a.weight + currentArc.distance))$
            **if not** $PQ.contains(a)$ **then**
                $PQ.insert(a)$
            **else**
                $PQ.decreaseKey(a, a.distance)$
            **end if**
        **end if**
    **end for**
    $RESULT.add(currentArc)$
**end while**
**return** $RESULT$
```

Concerning the technical details, each arc *a* has two real attributes: *weight* (set as inverse of "witness count") and *distance* (that always keeps the best found upper bound on the "aggregated distance" from *q* to *a*) as well as an *adjacentArcs* attribute that keeps the set of arcs sharing a node with *a* (except *a* itself). *PQ* is a min-type priority queue for keeping the arcs being processed, with the value of `distance` used as the priority, and *RESULT* is a set. *PQ* and *RESULT* are initially empty. We also assume that `infinity` is a special numeric value being greater than any real number.

More precisely, *PQ* is an *addressable priority queue* a standard extension of the priority queue abstract data structure. The interface of priority queue contains the following operations: *insert*() (for inserting new pair consisting of an element and its priority), *deleteMin*() (for efficiently identifying, returning and deleting a pair with currently the most urgent priority). In addition, the interface of an addressable priority queue contains the operation *decreaseKey*() that makes it possible to efficiently change (make lower) the priority of the element being its argument.

The simplest standard efficient implementation of an addressable priority queue is *binary heap* that makes it possible to achieve logarithmic time complexity for all the mentioned operations (if additionally the *merge*() operation, for merging two priority queues, is to be supported a more sophisticated data structure – *binomial heap* –

may be used to achieve logarithmic time complexity bounds). More precisely, each of the mentioned operations uses $O(log(n))$ comparisons of the element priorities, where $n$ is the number of elements in the structure. The reader can consult any classic textbook on basic algorithms, for example, Cormen et al. (1990) for more details concerning the Dijkstra algorithm or priority queues.

The output of the algorithm can be regarded as the selection of the top-$k$ most relevant (topologically closest) and important (based on arc weights) facts concerning the summarised entity. An example output computed on the base graph from Fig. 1 is presented on Fig. 2.

The algorithm is natural and efficient, but uses the same principles as the classic PRP ("Probability Ranking Principle") in IR, which has been criticised for risk of result redundancy. In other words, in the PRECIS'es output, the top-k most "relevant" and "important" facts in the summary may be very similar to each other. This is exactly what happens in the example on Fig. 2. The summary of "Tom Cruise" produced by diversity-oblivious PRECIS algorithm is dominated by "acted in" facts. Thus, a user interested, for example, in the private life of that entity would be dissatisfied by such output since, due to the over-representation of the "actedIn" arc labels, there is no fact concerning private aspects in the summary. See the discussion on the GES diversity and the example in the second half of Section 1.1.

Actually, the observed redundancy problem is not specific to the example on Fig. 2. We examined outputs of the PRECIS algorithm for about hundred other cases and noticed that arc label redundancy is a common problem with this algorithm.

Now we will propose a natural remedy for this problem. Namely, we introduce the *diversity* criterion (defined in Section 3.2) into the algorithm generating GES, in order to guarantee high coverage of arc labels in the output entity summary.

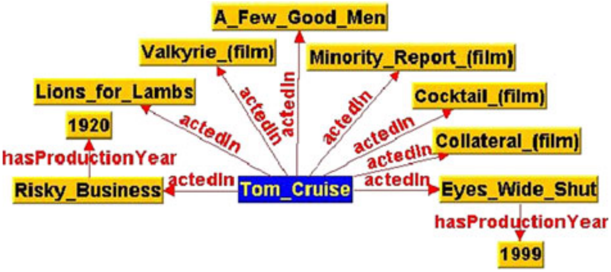
### 4.3 Diversity-aware entity summarisation algorithm DIVERSUM

We will now present another algorithm for the GES problem that takes into account the notion of *diversity* (Section 3.2), not only *relevance* and *importance*.

More precisely, the DIVERSUM algorithm will take into account the following, extended set of criteria while selecting the arcs (triples) to the summary:

- (first) *relevance* (arcs topologically close to the summarised node)
- (second) *diversity* (high coverage of different arc labels)
- (third) *popularity* (high arc label multiplicity)
- (fourth) *importance* (high "witness count", equivalent to low "distance")

**Fig. 2** Output of the PRECIS algorithm computed for entity "Tom Cruise" and $k = 10$ on the graph presented on Fig. 1. Notice high redundancy of labels of the selected arcs



The idea is as follows. The algorithm greedily selects triples from those connected to the current result (initially consisting solely of the summarised entity $q$), using the above four criteria in the presented order until collecting up to $k$ facts (arcs).

The pseudo code of DIVERSUM is shown in Algorithm 2.

**Algorithm 2** The DIVERSUM algorithm.

```
1:  $hops = 1$ (comment : *growing number of hops from q*)
2:  $S = \emptyset$ (comment : *the resulting summary, growing set of collected arcs*)
3:  $\Lambda = \emptyset$ (comment : *the growing set of collected unique arc labels*)
4:  **while** in $zone(hops, q)$ there is still an arc with the label that is not in $\Lambda$ **do**
5:      select a highest-multiplicity label $l \notin \Lambda$ in $zone(hops, q)$; $\Lambda.add(l)$
6:      among the arcs in $zone(hops, q)$ with label $l$ select the arc $a$ that has the minimum distance to $q$; $S.add(a)$
7:      **if** $S.size == k$ **or** $S$ already equals the input graph $D$ **then**
8:          return $S$
9:      **else**
10:         try to do next iteration of the while loop in line 4
11:     **end if**
12: **end while**
13: $hops$++; $reset\ \Lambda$; go to line 4
```

The description is as follows. The algorithm considers the arcs to be added to the summary in the order of *zones* that are increasingly distant from the summarised node. More precisely, $zone(i, q)$ (where $i \in N_+$) denotes the set of arcs in the distance of exactly $i$ hops from q. First, the algorithm considers only candidates from $zone(1, q)$ (i.e. the arcs incident to q) until the labels are exhausted, next, it focuses on $zone(2, q)$, $zone(3, q)$, etc. until the result consists of $k$ arcs or the underlying graph is exhausted. $\Lambda$ - stands for the growing set of unique labels present in $S$. Notice that in line 3 we *reset* the $\Lambda$ set. Different implementations of this operation allow for considering different variants of the algorithm. For example, if *reset* does nothing, we forbid arc label repetition in the whole summary; if *reset* makes the set empty, the uniqueness is forced only within each *zone*. We apply the latter variant in this article (i.e. an arc label can be repeated in other *zone*).

An example output of the algorithm is shown on Fig. 3, for the entity $q$ set to "Tom Cruise" and $k = 10$ on the input graph presented on Fig. 1. Notice that arc label "actedIn" is present twice. However, one instance is in $zone(1, q)$ and the other one in $zone(2, q)$. The same concerns arc labels "bornOnDate" and "livesIn".

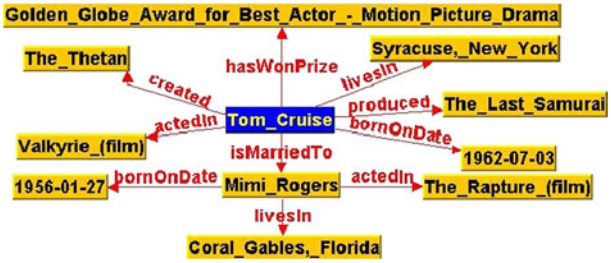


**Fig. 3** Output of the DIVERSUM algorithm for entity "Tom Cruise" with $k = 10$ computed for the graph depicted on Fig. 1. Notice the diversity of labels of selected arcs

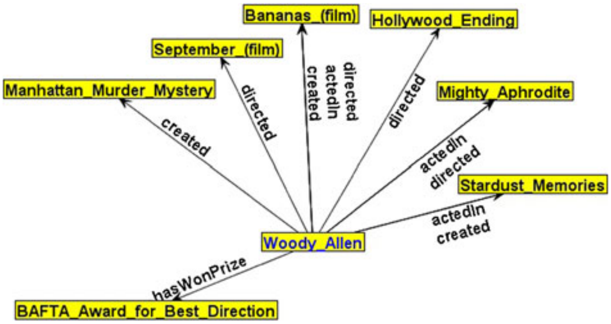


**Fig. 4** Small example of input used to illustrate how the algorithms work

One can easily notice that the result of DIVERSUM does not suffer from the redundancy problem that was present in the PRECIS output. This is due to the diversity-awareness of the algorithm. The example summary concerning "Tom Cruise" generated by the DIVERSUM algorithm covers multiple aspects of information concerning the entity due to its diversity, understood as high coverage of arc labels.

The experiment reported in Section 6 in a systematic way compares the arc label coverage of the DIVERSUM and PRECIS algorithms indicating that the former one almost always returns more diversified arc label set.

### 4.4 A toy example

In this section we will use a small example being a subgraph of the dataset extracted from IMDB database on movies to illustrate the difference between the two presented algorithms. The toy-example graph that will be used is presented on Fig. 4 and concerns some information about Woody Allen (represented by the central node) extracted from the IMDB knowledge base.

It contains 11 facts, represented as directed edges, about Woody Allen: 4 labelled as "directed", 3 labelled as "acted in", 3 labelled as "created" and 1 labelled as "has

**Table 1** List of arcs incident with the node "Woody Allen" in the graph shown in Figure 4 along with their multiplicities and distances (defined as inverse of "witness count" values that were available in the dataset) from this node.

| Subject | Predicate | Object | Label multiplicity | Distance |
|---|---|---|---|---|
| Woody Allen | directed | September (film) | 4 | $5.23E-6$ |
| Woody Allen | directed | Mighty Aphrodite | 4 | $5.29E-6$ |
| Woody Allen | directed | Bananas (film) | 4 | $6.13E-5$ |
| Woody Allen | directed | Hollywood Ending | 4 | $9.43E-5$ |
| Woody Allen | actedIn | Mighty Aphrodite | 3 | $5.29E-6$ |
| Woody Allen | actedIn | Stardust Memories | 3 | $1.76E-5$ |
| Woody Allen | actedIn | Bananas (film) | 3 | $6.13E-5$ |
| Woody Allen | created | Manhattan Murder Mystery | 3 | $5.74E-6$ |
| Woody Allen | created | Stardust Memories | 3 | $1.76E-5$ |
| Woody Allen | created | Bananas (film) | 3 | $6.13E-5$ |
| Woody Allen | hasWonPrize | BAFTA Award for Best Direction | 1 | $6.45E-3$ |

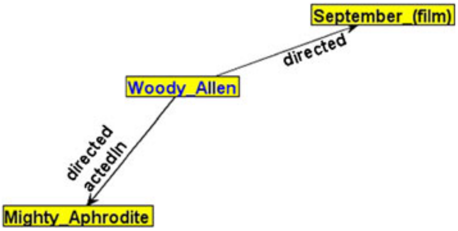


**Fig. 5** Output of PRECIS for input shown on Fig. 4 and presentation budget $k = 3$

won prize”. The arc weights (inverse of “witness count”) for this input graph are shown in Table 1.

Now, assume that we would like to obtain entity summary for Woody Allen with presentation budget limited to $k = 3$ facts (arcs) treating the small graph described above as the background knowledge database $D$.

PRECIS algorithm (described in Section 4.2) greedily selects arcs in the order from the closest to the farthest (to the summarised node) until presentation budget limit $k$ is reached. The result of the PRECIS algorithm for this case is presented on Fig. 5

The arcs are accepted in the following order (ties are broken arbitrarily):

1. Woody Allen - directed - September (film) (weight $5.23E - 6$)
2. Woody Allen - directed - Mighty Aphrodite (weight $5.29E - 6$)
3. Woody Allen - actedIn - Mighty Aphrodite (weight $5.29E - 6$)

Notice that two out of the three arcs selected by PRECIS have the same label.

DIVERSUM algorithm is designed to produce the output that is more diversified than PRECIS by avoiding arc label redundancy. More precisely, according to its description (Section 4.3) it first selects the label that is most popular among the arcs incident to the summarised node. In our case it is the label “directed” (multiplicity of 4). Out of the arcs labelled “directed” it selects the least “distant” from the summarised node (smallest arc weight). Then, it continues with next most popular arc labels and next least “distant” arcs labelled such. In other words, it selects the arcs in the lexicographic order based on label multiplicity and “distance” (ties are broken arbitrarily), such that no repetitions of arc labels are allowed in given zone.

The output of DIVERSUM in this case is shown on Fig. 6.

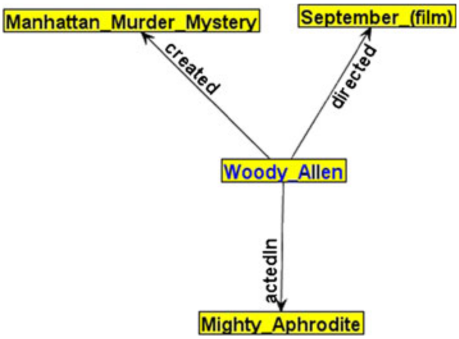


**Fig. 6** Output of DIVERSUM for input shown of Fig. 4 and presentation budget $k = 3$

The arcs are accepted in the following order:

1. Woody Allen - directed - September (film) (weight $5.23E-6$)
2. Woody Allen - actedIn - Mighty Aphrodite (weight $5.29E-6$)
3. Woody Allen - created - Manhattan Murder Mystery (weight $5.74E-6$)

If, in this example, the arc limit is $k=5$, then the next arc label to be selected by DIVERSUM is "Woody Allen hasWonPrize BAFTA Award for Best Direction" and then the algorithm starts looking for the next arcs to be selected in $zone(2, q)$, since the number of different arc labels in $zone(1, q)$ was only 4.

Even on this small example taken from real dataset one can observe the substantial difference of the two approaches. The second output consists of a more diverse set of facts.

## 5 Experimental setup

### 5.1 Implementation issues

Both algorithms: PRECIS and DIVERSUM have been implemented in Java programming language. It is important to notice that the algorithms concern only the problem of "which facts to select to the summary of a given entity?".

Other interesting problems like "how to layout the graphical results?" also had to be addressed but they are out of scope of this article.

The platform consists of many software components, including:

- data processing module
- fact selection summarisation algorithms (PRECIS and DIVERSUM)
- visualisation module
- graphical interface
- web-based module (for some experiments)

In this article we focus on the fact selection problem (PRECIS and DIVERSUM).

Importantly, we assume in this article that the data in the form of a semantic knowledge graph is already present and ready to use and contains verified information.

For visualising the results for this article we use the especially designed and implemented visualisation module that makes it possible to interact with the elements of the resulting entity summary by zooming, moving, hiding, etc. Our tool is based on the Jung library. For automatically computing the layout we use the novel AGNES algorithm (Sobczak et al. 2012). We also use a specially designed internal data format for efficient processing the underlying semantic graphs.

### 5.2 Dataset

A dataset for our experiments was extracted from the IMDB movie database,[7] and the YAGO ontology[8] containing information concerning 59,000 entities, including

[7] www.imdb.org

[8] http://www.mpi-inf.mpg.de/yago-naga/yago/

12,000 actors, over 530,000 edges representing 73 different binary relations. The "importance" weights on the arcs in this dataset were computed as inverse of, so called, "witness counts", (the number of times a textual representation of the given fact was found in a large corpus, see Elbassuoni et al. (2009) for details).

To make the results visually cleaner, we pruned from the base graph some relations of more technical nature, constituting redundant information or of little potential interest to the end users, so their omission would not affect the results significantly. The ignored relations are *type*, *hasImdb*, *hasLanguage*, *hasProductionLanguage*, *hasDuration*, *participatedIn*. This pre-filtering step was done after manual inspection of many results of both algorithms. In addition, some triples were pruned since they represent incorrect facts.[9]

### 5.3 Two levels of the arc limit value

We decided to evaluate the algorithms for two different levels of arc number limit to be presented in a summary: *low* (7 edges) and *high* (12 edges). It turned out that summaries with substantially more than 12 edges are not easy to comprehend in few seconds by humans (which is contradictory to the main function of a useful generic *summary* as described in the scenario in Section 1.1) and there was only a limited number of actors in the dataset for which both algorithms could generate much more than 12 edges. In addition, the value of 12 corresponds very well with the average number of facts present in the reference ground truth summaries used in one of our evaluation experiments (Section 7). Manual inspection shown that 7 edges are both easy to comprehend and is not too small, so that summary could contain different type of facts and is not too trivial. We also think that for many entities, especially famous, it would be very difficult to select significantly less than 7 facts to summarise even basic information about it.

### 5.4 Test sample

Concerning the choice of entities to be summarised in the experiments, we selected 20 prominent actors who starred in an over-average[10] number of movies , and for which both algorithms can produce at least 14 edges (facts) in the summarisation. The dataset contains also many other interesting types of entities to be potentially summarised, such as directors, for example. However, in order to obtain reliable results with limited resources, in the experiments we decided to focus only on a single type of entities being summarised. Since our dataset contains thousands of actors, we first pre-selected about 50 most active actors and then manually selected 20 out of it, considering the number of edges produceable in their summaries. The set was intentionally diversified in the terms of geographical and cultural context. More precisely, it included not only actors that are known to the global audience but also some very active actors known only in some particular regions of the world (e.g. some prominent Indian or European actors). This was done to test the summarisation algorithms in a wider range of levels of familiarity with the summarised entities.

[9] They exist due to the imperfect knowledge harvesting procedure, such as entity disambiguation.

[10] In the considered dataset the mean was about 3.17.

For the experiments reported in this article we computed outputs for all the 20 selected actors for both algorithms (PRECIS, DIVERSUM) and both levels of arc limit (k = 7 and k = 12). That resulted in 80 pictures. We used the software elements mentioned in Section 5.1 to produce this output sample.[11]

## 6 Experimental evaluation of diversity-awareness

In this section we objectively address the question whether DIVERSUM is actually more diversity-aware than PRECIS, that was hypothesised in Section 4.3.

To achieve this, we introduce a natural diversity measure for entity summaries that corresponds with the notion of diversity introduced in Section 3.2. We subsequently use this measure to systematically compare the PRECIS and DIVERSUM algorithms in terms of the level of diversity of the outputs they produce.

### 6.1 ALC: Arc-label-coverage diversity-aware evaluation measure

To objectively measure the level of diversity in a graphical entity summarisation, we introduce the following natural evaluation measure for the GES problem, called ALC (for "arc label coverage").

Given the graphical entity summary $S$ the ALC measure is defined as:

$$ALC(S) = |\{l : l \textit{ is an arc label in } S\}|$$

The ALC measure simply counts the number of different arc labels present in the summary what corresponds to the notion of diversity introduced in Section 3.2.

It is also possible to consider a "normalised" variant of the ALC measure:

$$NALC(S) = \frac{ALC(S)}{min(|S|, C_k)}$$

where $C_k$ is the number of different labels of arcs in the radius of $k$ (arc number limit) around the summarised node. The normalised variant of the measure is always upper bounded by 1 (maximum diversity), independently on arc limit value.

In this paper we use the simpler ALC variant of the measure.

It is important to notice that both variants of the measure satisfy the natural property of *monotonicity* defined in Sydow (2011) i.e. adding a fact to a summary cannot make the measure value lower for fixed presentation budget $k$.

### 6.2 Experimental results

To experimentally compare the diversity-awareness of the algorithms presented in Section 4, we computed the values of the ALC measure defined in Section 6.1 for all 80 graphical entity summaries constituting the test sample described in Section 5.4. The results are presented in Table 2

[11] The whole sample of produced outputs is available on e-mail request. In particular it can be used to repeat our experiments or conduct other.

**Table 2** The values of the ALC diversity measure, counting the number of different types of facts presented for each algorithm for k = 7 and k =12

| No. | Actor name | Precis 7 | Diversum 7 | Precis 12 | Diversum 12 |
|---|---|---|---|---|---|
| 1 | Ajay Devgan | 4 | **7** | 6 | **8** |
| 2 | Amitabh Bachchan | 4 | **6** | 5 | **6** |
| 3 | Anil Kapoor | 3 | **6** | 5 | **9** |
| 4 | Boris Karloff | 3 | **7** | 4 | **11** |
| 5 | Bruce Willis | 3 | **7** | 4 | **10** |
| 6 | Chevy Chase | 2 | **7** | 4 | **11** |
| 7 | Denzel Washington | 2 | **7** | 4 | **9** |
| 8 | Dharmendra | 6 | **7** | **9** | 8 |
| 9 | Gerard Depardieu | 5 | **7** | 6 | **10** |
| 10 | Henry Fonda | 3 | **7** | 4 | **11** |
| 11 | Jack Nicholson | 2 | **7** | 4 | **11** |
| 12 | John Wayne | 2 | **7** | 3 | **12** |
| 13 | Laurence Olivier | 3 | **7** | 6 | **11** |
| 14 | Marlon Brando | 5 | **7** | 7 | **12** |
| 15 | Richard Gere | 2 | **7** | 4 | **10** |
| 16 | Robert De Niro | 4 | **7** | 6 | **12** |
| 17 | Robert Mitchum | 3 | **7** | 4 | **10** |
| 18 | Steven Seagal | 4 | **7** | 5 | **11** |
| 19 | Sylvester Stallone | 4 | **7** | 4 | **10** |
| 20 | Tom Hanks | 3 | **7** | 4 | **12** |
|  | Average: | 3.4 | 6.85 | 4.85 | 10.3 |

Value in bold shows which algorithm achieved higher ALC value

The result is very clear. The following observations can be made:

- DIVERSITY beats PRECIS in all cases for k = 7 and in 95 % cases for k = 12
- in terms of the averaged values, DIVERSITY definitely beats PRECIS, achieving over twice higher arc label coverage than PRECIS
- the higher the arc limit, the higher the average arc label coverage

To summarise, on the tested sample, DIVERSUM provides definitely more diverse set of arc labels in the generated entity summarisation than PRECIS.

## 7 Fact selection evaluation experiment

In this section, we evaluate in an objective and systematic manner the graphical summarisations generated by the PRECIS and DIVERSUM algorithms in terms of important fact coverage.

This is achieved by an intrinsic evaluation experiment inspired by the ROUGE evaluation method commonly used for textual summarisation evaluation (Lin and Hovy 2003).

The original ROUGE method is based on a reference set of ground truth text summaries that are prepared for each textual document in the test sample. The tested summarisation algorithm is evaluated by computing the recall of the summarisation units (usually sentences) i.e. the proportion of the summarisation units produced by the algorithm to the total number of units present in the ground truth summary.

Analogously, in our problem of evaluation of two algorithms for the GES problem, for each of the 20 entities (actors) from the test sample (Section 5.4) we prepared a reference, ground truth summary that consists of some facts concerning the entity. More precisely, the reference summaries were extracted from the Wikipedia info-boxes concerning the summarised entities. Because Wikipedia is a portal that is edited by a large population of editors and has some natural quality-control mechanisms, it can be assumed that info-boxes in Wikipedia, concerning entities can be viewed as high-quality summaries concerning the entities. Also their sizes fit very well the arc limit values of 7 and 12. The number of facts concerning an entity in an info-box ranged from 5 to 21, depending on the entity, with the average value of 12.5.

Next, we compared each of the 80 test sample (Section 5.4) summaries with the corresponding reference summaries in terms of fact selection. More precisely, we computed how many facts from the info-box reference summary are "hit" by the facts from the evaluated summaries.

Because the facts concerning entities may be presented in various forms, in particular they can be expressed more or less generally, we designed and applied some logical rules to judge whether there was a "hit" or not.

The main, most general rule was as follows. We counted the "hit" whenever a fact in the reference summary could be directly logically implied by a fact or facts in "our" summary (or, equivalently if "our" fact is more specific or identical). For example, a presence of a fact "acted in" in "our" summary implies the fact "the summarised entity is an actor", etc. Having the number of "hits" for each evaluated summary we computed the recall values. The results are presented in Table 3.

The results of the experiment definitely indicate the superiority of the diversity-aware algorithm (DIVERSUM) over the diversity-oblivious one (PRECIS) on this test sample. In particular, DIVERSUM beats PRECIS by far in terms of correctly covering facts from the reference Wikipedia summaries in 90 % of cases for k = 7 and in 95 % for k = 12. Notably, the performance of DIVERSUM is *never* worse than that of PRECIS. Also, the average performance figures (last row in Table 3) show that DIVERSUM covers twice more facts than PRECIS on average.

One may draw a conclusion that the higher degree of diversity-awareness of the DIVERSUM algorithm (that was actually objectively confirmed by the experiment reported in Section 6) makes it better selects important facts for generic summaries of entities, on the considered test sample.

Importantly, one can easily observe that the performance of both algorithms generally grows with the value of $k$ (presentation budget) which seems to be a desired property of any reasonable summarising algorithm.

Regarding the absolute recall values achieved by the algorithms, they depend strongly on the entity, ranging from around 6 % (very poor) for some rare cases to over 60 % (quite good) with an average value of over 40 % for DIVERSUM with k = 12. Although it might seem that there is a lot of room for improvement in terms of recall values achieved by the algorithms, there are some reasons for regarding these figures as quite promising.

First of all, the average recall figures are lowered by the cutting value of $k$ (arc limit) for cases where the reference summary contains more than $k$ facts.

Second, we observed that many facts in the outputs of DIVERSUM that were more than one hop from the summarised node where actually present very early in the full Wikipedia entry while not present in the info-boxes.

**Table 3** Fact coverage in respect to Wikipedia entry for each entity for both algorithms for k = 7 and k = 12

| Actor Name | Precis 7 | | Diversum 7 | | Precis 12 | | Diversum 12 | | #Facts |
|---|---|---|---|---|---|---|---|---|---|
| | # Hits | Recall | # Hits | Recall | # Hits | Recall | # Hits | Recall | |
| Ajay Devgan | 3 | 30 | **5** | **50** | 5 | 50 | **6** | **60** | 10 |
| Amitabh Bachchan | 3 | 30 | **4** | **40** | 3 | 30 | **6** | **60** | 10 |
| Anil Kapoor | 2 | 16.67 | **5** | **41.67** | 2 | 16.67 | **5** | **41.67** | 12 |
| Boris Karloff | 1 | 12.50 | **3** | **37.50** | 1 | 12.50 | **3** | **37.50** | 8 |
| Bruce Willis | 3 | 37.50 | **5** | **62.50** | 3 | 37.50 | **5** | **62.50** | 8 |
| Chevy Chase | 1 | 5.88 | **2** | **11.76** | 1 | 5.88 | **2** | **11.76** | 17 |
| Denzel Washington | 1 | 20 | **3** | **60** | 1 | 20 | **3** | **60** | 5 |
| Dharmendra | 3 | 23.08 | **6** | **46.15** | 4 | 30.77 | **7** | **53.85** | 13 |
| Gerard Depardieu | 1 | 12.50 | **3** | **37.50** | 2 | 25 | **3** | **37.50** | 8 |
| Henry Fonda | 1 | 6.25 | **5** | **31.25** | 1 | 6.25 | **5** | **31.25** | 16 |
| Jack Nicholson | 2 | 14.29 | **6** | **42.86** | 2 | 14.29 | **6** | **42.86** | 14 |
| John Wayne | 1 | 5.88 | **5** | **29.41** | 1 | 5.88 | **7** | **41.18** | 17 |
| Laurence Olivier | 3 | 21.43 | **5** | **35.71** | 3 | 21.43 | **7** | **50** | 14 |
| Marlon Brando | 2 | 9.52 | **4** | **19.05** | 3 | 14.29 | **7** | **33.33** | 21 |
| Richard Gere | **2** | **14.29** | 2 | 14.29 | **2** | **14.29** | 2 | 14.29 | 14 |
| Robert De Niro | 3 | 18.75 | **4** | **25** | 3 | 18.75 | **4** | **25** | 16 |
| Robert Mitchum | 2 | 25 | **5** | **50** | 2 | 25 | **5** | **50** | 8 |
| Steven Seagal | **4** | **23.53** | 4 | 23.53 | 4 | 23.53 | **5** | **29.41** | 17 |
| Sylvester Stallone | 3 | 23.08 | **5** | **38.46** | 3 | 23.08 | **6** | **46.15** | 13 |
| Tom Hanks | 2 | 22.22 | **5** | **55.5** | 3 | 33.33 | **6** | **66.67** | 9 |
| Average | 2.15 | 18.62 | 4.25 | 37.61 | 2.45 | 21.41 | 4.95 | 42.75 | 12.5 |

Values in bold denote whether the DIVERSUM beats PRECIS for k = 7, and for k = 12. The last three column contains the total number of facts in the corresponding ground truth summary. The values in all the "Recall" columns represent percentage. The last row contains average values in the corresponding columns

Finally, as it is frequently reported in the area of text summarisation, usually there is no consistency between reference summaries generated by different methods, different experts or even by the same expert in different moments. Thus, achieving recall of over 40 % for a *single* generic reference summary seems to be a quite promising result.

## 8 Expert-based assessment experiment

Apart from the objective experimental evaluation of fact selection that was reported in the previous section, we conducted additional, user-based intrinsic experiments, in which graphical entity summaries (in the form of graphs) were presented to human evaluators to be subjectively assessed as generic summaries.

In particular, the experiment aim was to:

- test whether outputs of diversity-aware algorithm (DIVERSUM) are significantly preferred by the human evaluators as generic summaries of information concerning the entities without any specified task
- collect valuable comments and user feedback, concerning both algorithms, that can serve for improving the summarisation algorithm in future work.

The positive answer to the first question would be additional evidence of superiority of diversity-aware DIVERSUM over the PRECIS algorithm.

The idea of the user-based experimental evaluation was simple: to repeatedly present human evaluators with a pair of outputs produced by PRECIS and DIVERSUM (in random order), next to each other, concerning *the same entity* and the same number of presented facts.

The users responded to the presented output pairs by putting comments on them and answering simple questions on, for example:

- which output is preferred and why
- what is evaluators' familiarity with the entity
- how many interesting facts they found in the summary
- whether they miss any expected facts, etc.

Before the experiment, the evaluators were given precise instructions, including a short description of the generic summarisation problem. To avoid any bias, the evaluators were not told any details about the summarising algorithms, most importantly, they were *not informed* that diversification was the issue studied in the experiment.

In addition, the evaluators were asked to focus more on the *choice of facts* in the summary rather than the particular graphical layout of the resulting graph.

In this way we tried to minimise the impact of the particular *visualisation* algorithm used in this experiment. However, by showing the summaries in the graphical form we made it possible to collect potentially valuable feedback on the graphical aspects too, in the form of open-text comments.

## 8.1 Implementation of the experiment

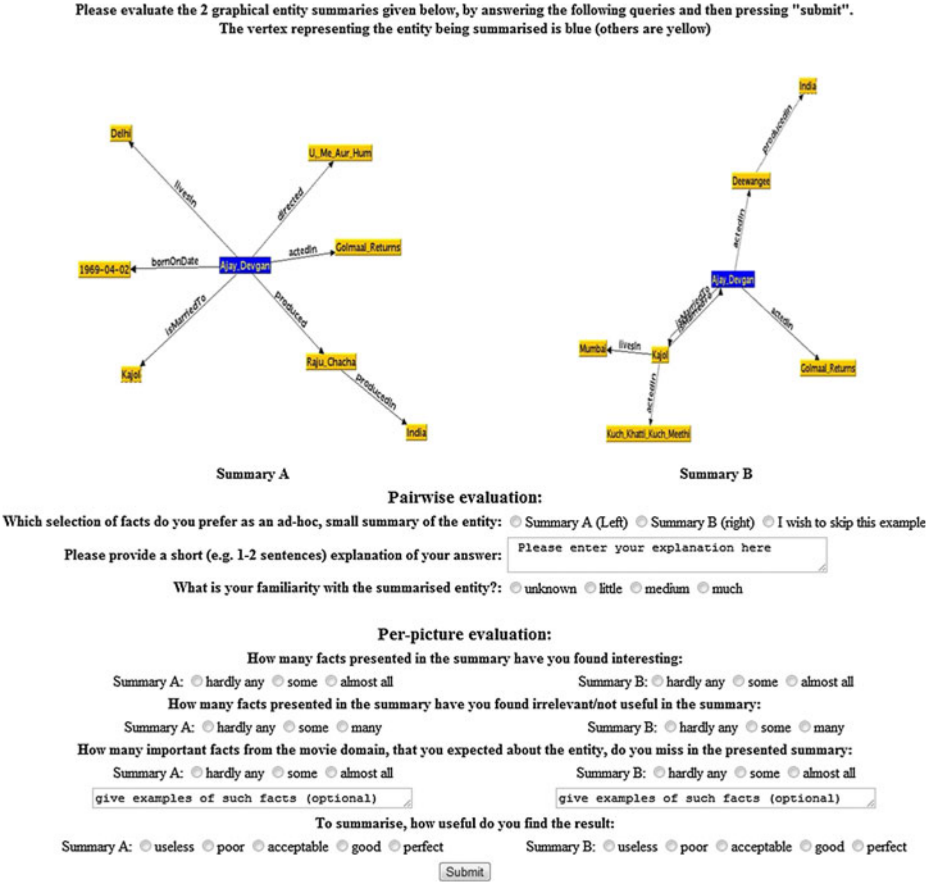
We designed and implemented a web application for presenting the pairs of outputs of the algorithms and providing a form for collecting the user answers and open-text explanations for their the choice. The evaluation form is presented on Fig. 7.

Next, we invited (by an e-mail announcement) a pool of anonymous volunteer evaluators being the researchers from the "Databases and Information Systems" Department of MPII, Saarbruecken, that were not involved in or even aware of the project, so that their objectivity and honesty could be assumed.

The evaluators asynchronously visited the web page to assess the generated outputs.

Although, due to privacy protection no personal information concerning the evaluators was recorded, it is worth mentioning that the population from which the evaluators where chosen (i.e. the members of the mentioned institute department) represent extreme diversity in terms of cultural background (Europe, Asia, Africa) and could be regarded as experts in the computer science domain and especially used to graph-structured data. The age ranges approximately from 25 to 35, the gender is diverse. The familiarity with the movie domain was one of the questions and the collected data clearly shows that it was also very diverse, ranging from unfamiliarity to high familiarity (see Fig. 10, left).

In our opinion the diversity among the evaluation team fits well with assessing summaries for an unknown information need.



**Please evaluate the 2 graphical entity summaries given below, by answering the following queries and then pressing "submit".**
**The vertex representing the entity being summarised is blue (others are yellow)**

**Pairwise evaluation:**

**Which selection of facts do you prefer as an ad-hoc, small summary of the entity:** ○ Summary A (Left) ○ Summary B (right) ○ I wish to skip this example

**Please provide a short (e.g. 1-2 sentences) explanation of your answer:** Please enter your explanation here

**What is your familiarity with the summarised entity?:** ○ unknown ○ little ○ medium ○ much

**Per-picture evaluation:**

**How many facts presented in the summary have you found interesting:**

Summary A: ○ hardly any ○ some ○ almost all Summary B: ○ hardly any ○ some ○ almost all

**How many facts presented in the summary have you found irrelevant/not useful in the summary:**

Summary A: ○ hardly any ○ some ○ many Summary B: ○ hardly any ○ some ○ many

**How many important facts from the movie domain, that you expected about the entity, do you miss in the presented summary:**

Summary A: ○ hardly any ○ some ○ almost all Summary B: ○ hardly any ○ some ○ almost all

give examples of such facts (optional) give examples of such facts (optional)

**To summarise, how useful do you find the result:**

Summary A: ○ useless ○ poor ○ acceptable ○ good ○ perfect Summary B: ○ useless ○ poor ○ acceptable ○ good ○ perfect

**Fig. 7** Evaluation form presented to users via a web server in the intrinsic assessment experiment

From subsequent analysis of access logs it seemed that the number of actual active evaluators was between 10 and 20. The application was designed so that it was very unlikely that the same user would see the same pair of outputs more than once (in such case, the evaluator was instructed to skip the example).

## 8.2 Preference of DIVERSUM

We collected 71 assessments within three days, out of which 66 were complete. Each of the 20 selected actors received at least 2 assessments (18 actors received more) with median 3, mean 3.3, and maximum of 6 assessments.

It turned out that the users expressed quite strong preference. In over 80 % of cases, the output of DIVERSUM was preferred to the other one. Remind that the outputs where presented in random order without any additional identification information. The result is statistically significant with the value p = 3E-7 according to sign test (Wackerly et al. 1990). DIVERSUM was even stronger preferred for the smaller value of arc number limit $k = 7$ i.e. in over 90 % of cases (Fig. 8).

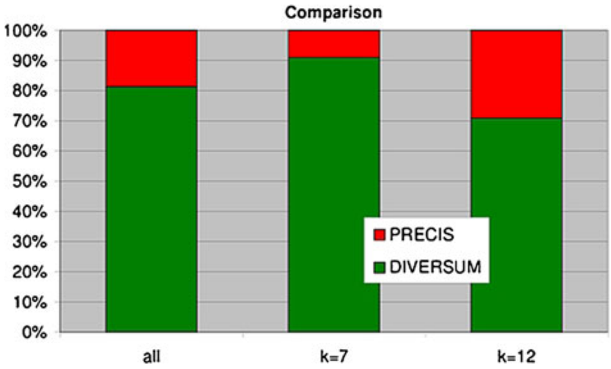


**Fig. 8** Fraction of cases where human evaluators preferred DIVERSUM to PRECIS

Thus, the first aim listed in the beginning of the Section 8 was clearly addressed: the outputs of diversity-aware algorithm was significantly preferred by the evaluators in this experiment.

## 8.3 The role of diversity in the results

One may claim that what was actually assessed in the experiment was some particular diversity-aware *algorithm* not the notion of diversity in the summaries itself.

While such observation is generally reasonable, we have an argument that the notion of diversity actually *was* the reason of the preference to large extent.

The evidence for this was found in open-text comments given by the evaluators. Namely, it turned out that in numerous cases evaluators indicated the notion of "diversity" as the actual reason for the preference.

Such evidence is quite strong because the evaluators where not directly asked about diversity in any way nor were aware that the project concerns the notion of diversity.

Precisely, 41 (out of 66 valid) assessments contained open-text explanation for the preference. 39 % of them (i.e. 16 comments) explicitly explained the preference of DIVERSUM using the word "diversity" or "diverse", etc. For example, *"Summary A has important and diverse relations summarising life and achievements of the actor", "Facts with diverse relations in Summary A", "There are only actedIn relationships in the summary B. Summary A is more diversified - it has a lot more information"* . All the user comments of this kind are presented in the Appendix at the end of the article.

In addition, out of the remaining 25 comments, 9 contained equivalent explanation (i.e. appreciating broader aspect coverage) however expressed without the word "diversity". For example: *"covers more topics, not just actedIn", "more types of facts"*, etc. All comments of this kind are presented in the Appendix.

In total, in 60.9 % of commented cases the preference of DIVERSUM was directly or indirectly explained by the diversity-awareness in terms of arc label coverage.

To complete the picture, we admit that 1 comment exceptionally suggested the preference of *homogeneity* in the summary: *"I get more better image in my mind about the actor with the summary of the films he has worked in or created from Summary B rather than his bio data from Summary A."*.

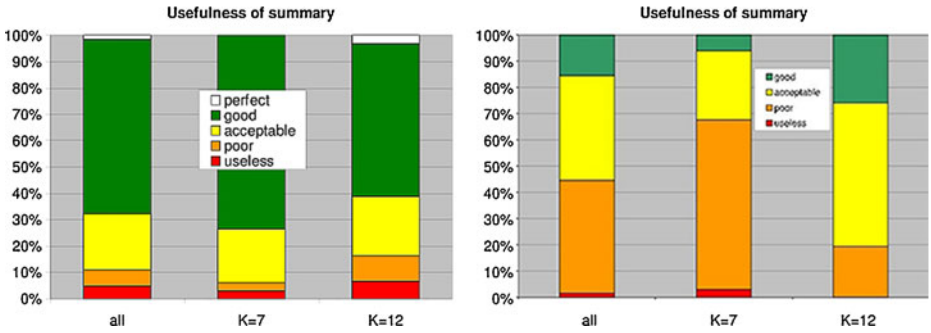


**Fig. 9** Feedback on outputs of the compared algorithms viewed as general summaries concerning the entities. *Left*: DIVERSUM, *Right*: PRECIS

Another important groups of comments concerned: "better readability" (2 comments), higher "relevance" of facts (2 comments), "importance" or "informativeness" of facts (2 comments). Two comments regretted for "too long paths" in DIVERSUM.

### 8.4 Additional observations

Concerning the assessment of the algorithms' outputs viewed as generic summaries of the entities the users could select for each of the compared outputs one of the following marks: "perfect", "good", "acceptable", "poor", "useless". The median mark for DIVERSUM was "good". Only 5 % of cases it was marked as "poor" or "useless" for the low level of limit budget ($k = 7$) and in about 10 % in total. Diversum summary was once marked as "perfect". The PRECIS algorithm received remarkably poorer marks with median being "acceptable". More details on Fig. 9.

Concerning the familiarity of the users with the summarised entities in about 50% of the cases the familiarity was marked as "high" or "medium" and in other as "little" or even "unknown" (see the left part of Fig. 10).

DIVERSUM was appreciated by the users, across the full range of familiarity (see the right part of Fig. 10). In particular, for high familiarity it was preferred in 83 %

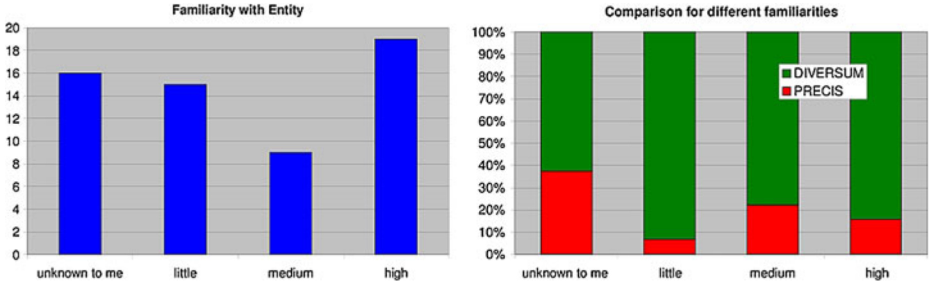


**Fig. 10** *Left*: familiarity of the evaluators with the summarised entities (number of cases); *right*: preference between DIVERSUM and PRECIS for different levels of familiarity

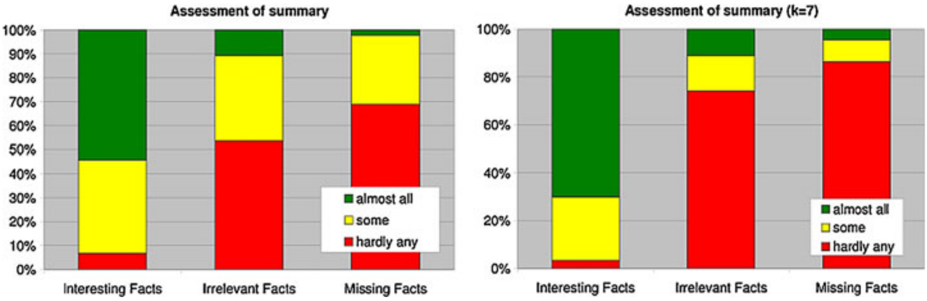


**Fig. 11** DIVERSUM: Comparison of assessment of selected facts for different edge limits: $k = 12$ (*left*) and $k = 7$ (*right*)

cases, for little in over 90 % cases. The algorithm seems to perform a bit worse for unknown entities, but it is still preferred here in about 62 % of cases.

### 8.5 Non-comparative assessment of the algorithms

Considering more direct assessments of the facts selected by the DIVERSUM algorithm, the results are also very promising: in over 93 % of cases the summary was assessed as having “almost all” or “some” interesting facts, and in about 98 % of cases the evaluators did not complain that the summary was missing most important facts. In only about 11 % of cases the summaries contained facts assessed as “irrelevant” by evaluators. See Fig. 11 for details.

There is a significant difference of user experience concerning the choice of facts selected by the DIVERSUM algorithm for two different edge limits (Fig. 11). Similarly to the previously reported aspects of summaries, the algorithm performs better for low edge limits (except the assessment of the “missing facts” is obviously performing better for higher edge limit).

Overall, one observes that the diversified entity summarisation algorithm received definitely better assessments than the diversity-oblivious baseline in all asked aspects. It is clear that across all measured properties the attractiveness of the diversified summarisation is higher for the smaller summarisation size limit. It has an intuitive explanation: as the number of possible facts to select decreases, the algorithm should pay more attention to try to cover more diversified aspects of the entity.

## 9 Crowdsourcing-based assessment experiments

This section reports an additional user-based experiment that completes the assessments presented in this article. It is similar to the experiment presented in Section 8 in terms of its goals and general idea but differs in one important element.

While the previous one was based on evaluators that could be regarded as experts in the field of information systems we decided to repeat it on a broader audience of anonymous web users. This is because of the potential of using GES as generic entity summarisation by a very broad population of users (e.g. as a feature in next-generation web search engines, for example).

To simulate the anonymous, broad audience, we adapted the previous experiment to the *crowdsourcing* approach.

After a short introduction to this technique in Section 9.1 we present the setup of both experiments in Section 9.2, followed by the result presentation and discussion in Section 9.3.

### 9.1 The idea of crowdsourcing approach

Crowdsourcing (Howe 2008) is a relatively new approach. It refers to delegating tasks to a group of people or community (crowd) through an open call (Kittur et al. 2008). The idea behind this concept is to split one big task into a set of repetitive singular tasks that can be done independently. Every contributor that completes such task is awarded with a small amount of money. Those singular tasks can be done in parallel by many contributors. Since some crowdsourcing services offer over 1 million individual contributors, the speed advantage of this approach is obvious. Importantly, the contributors represent a broad sample of random Internet users with different backgrounds, not being a group of experts in any particular area. They form a valuable target group for user evaluation that ideally complements the expert group used in our previous experiment. This concept is already being research as a potential new area for problem solving conducting evaluation experiments (Kittur et al. 2008; Brabham 2008; Alonso et al. 2008)

### 9.2 Setup of the crowdsourcing experiments

Because crowdsourcing is a relatively new service and only limited functionality is provided in currently available platforms, our expert-based experiment setup could not be directly applied due to technical restrictions. After additional research we picked an appropriate service provider – Crowd Flower – a crowdsourcing service with over 1.5 M workers which uses Amazon Mechanical Turk.[12] Anyway, no available crowdsourcing platform offered the full functionality needed in our experiments, in particular, it concerned displaying graphical output of our algorithms. Thus, some modifications were needed in order to meet our constraints. Those changes were of a technical nature and did not affect the question content. We had to split the survey form so that only the part of it was placed on the crowdsourcing service. The other part was deployed on our own web server and provided screenshots of two graph extracts which were part of the initial evaluation form as presented on Fig. 7. The appropriate link to the page with screenshots was added to the form deployed on the Crowd Flower service to link the parts. We also enriched the form with one additional field. The contributor was supposed to input into that field the unique ID of the screenshot set presented to him when he followed the link provided in the form. This allowed us to connect a particular screenshot with the given answer set and was used as simple verification tool to distinguish valid results from invalid ones.

[12] We could not use Mechanical Turk directly due to restrictions placed by Amazon that allowed only US residents to publish the tasks.

If the answer set lacked this ID or the ID was wrong, the whole questionnaire was regarded as invalid and excluded from further analysis.

### *9.2.1 First crowdsourcing experiment*

The experiment consisted of 200 individual assessments, each one concerning showing a pair of outputs to be completed by a contributor and being awarded by a small amount of money according to the crowdsourcing service model. Thanks to large number of contributors available trough Crowd Flower all 200 surveys were completed in around 24 hours. 45 out of 200 samples were rejected due to a missing or incorrect ID value. Such invalid entries could have been caused by a contributor not understanding the task description or trying to dishonestly maximise financial benefit without paying attention to the quality of the delivered work.

### *9.2.2 Second crowdsourcing experiment with improved quality control*

To increase the reliability of the crowdsourcing evaluation, we repeated the experiment with additional elements of quality control. In order to eliminate dishonest contributors (Kittur et al. 2008) interested only in financial benefits, not in the quality of the delivered work, we applied two techniques: forcing increased time involvement of the evaluators and test questions. Technically, two modifications of the answer form were made to achieve these goals. We changed all radio buttons to text input fields to force the user to manually enter appropriate answers instead of quickly selecting a radio button. The second modification was introduced to test whether a contributor understands the information presented in the graph extract. An additional field was added where the user was asked to type in the name of the summarised entity. This served as an additional validation test. Everything else remained unchanged.

As the result we achieved a significant 20 % drop in the number of invalid answers (44 vs 55 out of 200).[13] The change resulted also in 70 % of increase in time spent by a single user on completing the task (6 minutes on average compared to about 3.5 minutes in the first experiment) that could cause higher quality of answers.

## 9.3 Preference of DIVERSUM

In both crowdsourcing experiments the preference of DIVERSUM over PRECIS was similar: almost 60 % to 40 % (see Fig. 12). The results are statistically significant with $p = 0.044$ and $p = 0.084$ according to sign test (Wackerly et al. 1990), respectively. These results further support the results presented in Sections 7 and 8.

## 9.4 Users' textual comments

While the participants of the crowd-sourcing experiments rarely (roughly in 25 % of cases) provided textual comments on their forms, some comments were very valuable as they again support the general intuition that covering diversified set of aspects is a desired property of an algorithm for the GES problem.

[13] This improvement could have been even better but misspellings and other errors created by the usage of an input text field contributed to a lower gain then expected initially.

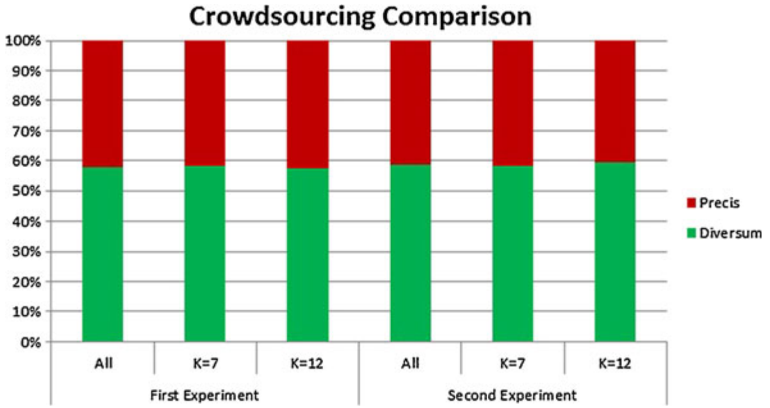


**Fig. 12** Comparison of user preferences based on results from first (*left*) and second (*right*) crowdsourcing experiment

More precisely, about half of valid comments explicitly appreciated more diversified coverage of facts concerning various aspects of the entities that was provided by the DIVERSUM algorithm. Additionally, some comments demonstrated how diverse and unexpected can be user preferences. A good example of a user information need that is hard to predict is the following comment: *"This algorithm shows the family concept. It is nice"*). In our opinion, such comments provide an additional argument for the need of diverse summaries in the context of unknown user interests.

On the other hand, very few comments (about 3 in total) expressed the opposite preference, i.e. the point of view that, for example, a summary of an actor should contain mostly the information about movies they acted in.

The most interesting (also negative) user comments are listed in the Appendix.

The remaining part of other valid comments seemed too general to be useful to draw any concrete conclusions (e.g. "better readability", etc.).

## 9.5 Additional results

In this subsection we present additional results per analogy to Section 8 to complete the picture. In general, these results are compatible with those presented in Section 8.

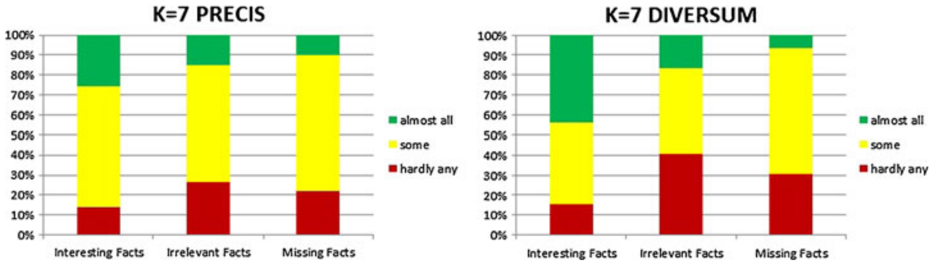


**Fig. 13** Comparison of assessment of selected facts for PRECIS (*left*) and DIVERSUM (*right*) on edge limit: $k = 7$ for the first crowdsourcing experiment

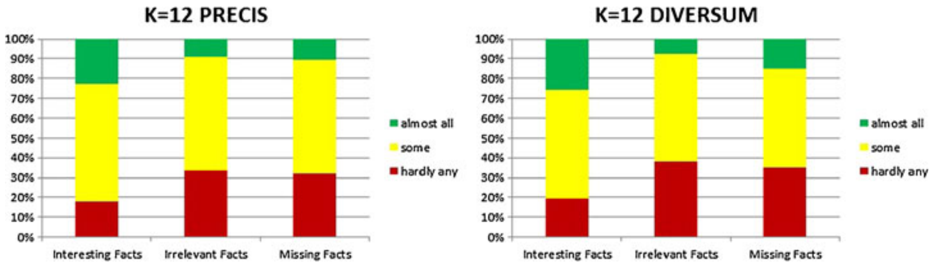


**Fig. 14** Comparison of assessment of selected facts for PRECIS (*left*) and DIVERSUM (*right*) with edge limit $k = 12$ for the first crowdsourcing experiment

In the first experiment Section 9.2.1, as Figs. 13 and 14 show DIVERSUM provided better results according to evaluators. For both edge limits DIVERSUM achieved better performance than PRECIS for almost all possible scores. Moreover, for $k = 7$, almost twice as much users (44 %) as in the case of PRECIS (25 %) claimed that DIVERSUM displays ”almost all” interesting facts about the presented entity. Similar patterns concern “irrelevant fact” and “missing fact” categories (Fig. 15).

This shows that users appreciate the diversified query result especially when the number of elements that can be displayed as the query result is limited.

Very similar results were obtained for the second experiment (Section 9.2.2). Figure 16 shows that much more users than in the first experiment pointed out that DIVERSUM contains ”almost all” interesting facts (more than the PRECIS). On the other hand the same user group pointed out that diversified variant has more ”almost all” missing facts then PRECIS one. This seems to show that users can have different perception of the presented summary up to the point of results being contradicted. Fig. 17, 18 also show that the diversified result for $k = 12$ has been even more favoured in this experiment than in the previous one.

Importantly, for $k = 12$ around 60 % of users ranked it as ”perfect” or ”good” in terms of usefulness, which is even more than for $k = 7$. Furthermore, the results obtained for the crowdsourcing experiments show that more contributors ranked results as being ”perfect” than in case of evaluation made by experts 8.

Finally, Fig. 19 presents general user preferences (combined data for both $k = 7$ and $k = 12$) against user familiarity with the movie domain. It indicates a general trend - the more knowledge of the domain users have, the higher is their preference

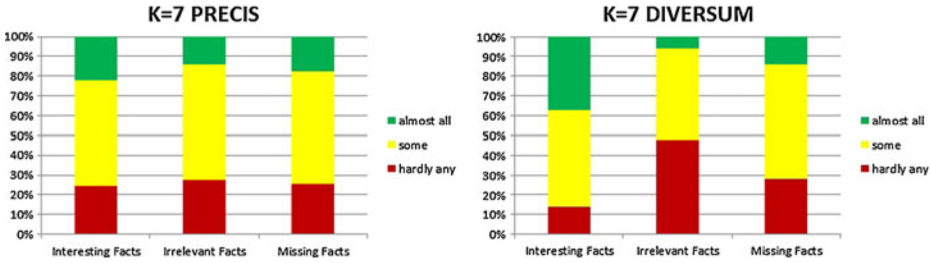


**Fig. 15** Comparison of assessment of selected facts for PRECIS (*left*) and DIVERSUM (*right*) on edge limit: $k = 7$ for the second crowdsourcing experiment

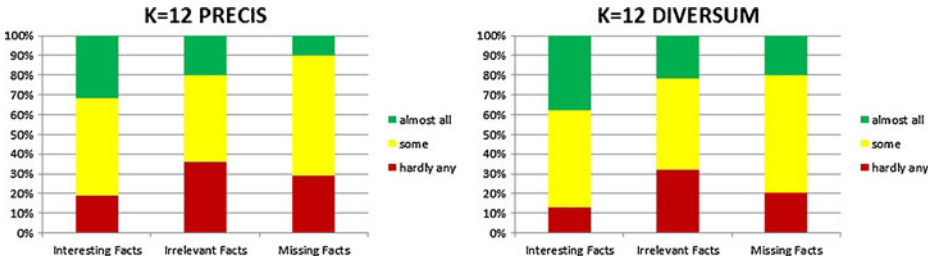


**Fig. 16** Comparison of assessment of selected facts for PRECIS (*left*) and DIVERSUM (*right*) with edge limit $k = 12$ for the second crowdsourcing experiment

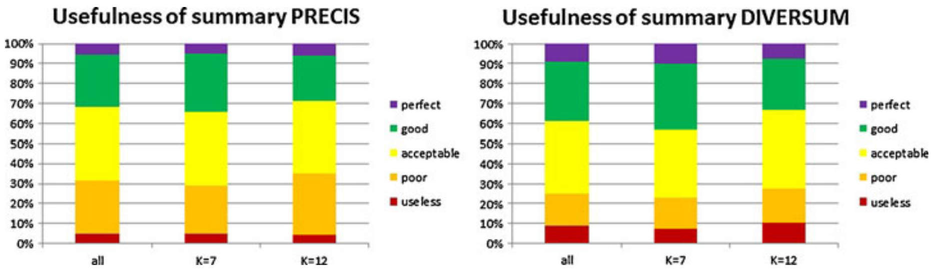


**Fig. 17** Comparison of usefulness of selected facts for PRECIS (*left*) and DIVERSUM (*right*) for the first crowdsourcing experiment

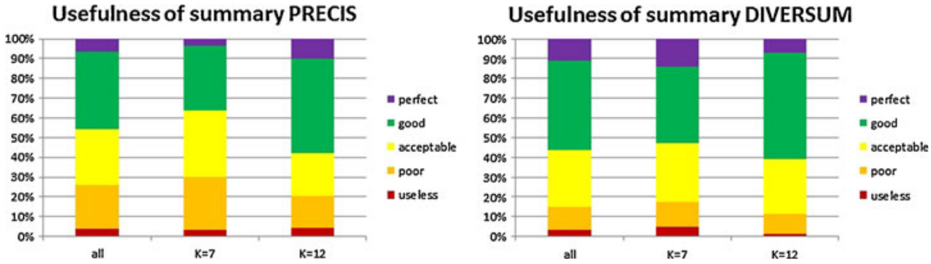


**Fig. 18** Comparison of usefulness of selected facts for PRECIS (*left*) and DIVERSUM (*right*) for the second crowdsourcing experiment

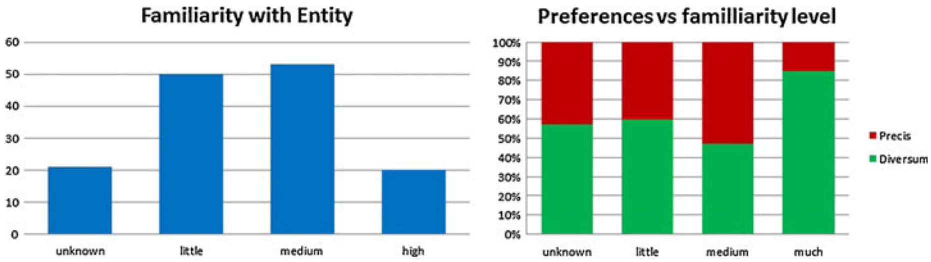


**Fig. 19** *Left*: Crowd Flower community familiarity with movie domain (number of cases); *right*: preference between DIVERSUM and PRECIS for different levels of familiarity. Values for first crowdsourcing experiment

of DIVERSUM vs PRECIS. An exception is the “medium” level, where DIVER-

SUM and PRECIS achieved similar figures (48 % vs 52 %). In particular, about 84 % of the users who declared high familiarity with the domain preferred diversified summaries. This result is compatible with the one obtained for the expert users (Section 8).

## 10 Conclusions, discussion and further work

This article concerns the role of diversity in a quite novel problem of graphical entity summarisation. After giving motivations for studying such problem and short overview of the related text summarisation problem and the notion of diversity in information retrieval the ideas on how to adapt it to the new problem are proposed.

A baseline algorithm PRECIS for GES is presented together with observations on lack of diversity in its output. Then, a diversity-aware measure, called ALC (for “arc label coverage”), for GES is proposed and another algorithm, called DIVERSUM, that is especially designed to be diversity-aware is presented. As is demonstrated on a sample from a real dataset, DIVERSUM is indeed more diversity-aware in terms of the ALC diversity measure.

We report an intrinsic experiment that evaluates the quality of fact selection of both algorithms. The result of this experiment clearly indicates superiority of diversity-aware DIVERSUM over the diversity-oblivious PRECIS in terms of achieving higher recall on ground truth reference entity summaries extracted from Wikipedia.

In addition to the above, we report another experimental result where human experts subjectively assess DIVERSUM as significantly better algorithm for generic graph entity summarisation than PRECIS.

In the applied methodology it was necessary to make some simplifying assumptions. One of the most important ones was that the abstract concept of diversity in our experiments was represented by a single (DIVERSUM) algorithm and one might argue that the user assessments may be biased by the particular choice of algorithm or visualisation method.

However, the fact that many users explicitly and spontaneously commented their preference (of DIVERSUM over PRECIS) by the higher diversity of the summary seems to clearly indicate that the notion of diversity was the reason of preference, to much extent. Remind that users were not informed that the experiments considered diversity.

Similar results were obtained in additional two reported experiments that are repeated on a sample of a broad, anonymous population of Internet users by means of a crowd-sourcing approach.

Furthermore, multiple collected user comments demonstrate how different are user information needs that additionally implies the need for diversity.

On the other hand, designing an experiment that tests the preference of pure diversity as a concept independently on the particular algorithm or visualisation method seems to be a challenging task and may be an interesting follow-up of the work presented in this article.

To summarise, all the results of reported objective and subjective experiments are compatible and indicate the preference of the diversity-aware DIVERSUM algorithm over the diversity-oblivious one, and that diversity-awareness is a needed property in the GES problem.

### 10.1 Future work

We observed the following problems while analysing the results in all the reported experiments:

- some results seem to suffer a "topic drift" i.e. presenting some facts that seem to be too distant from the entity summarised
- sometimes, the algorithm selects irrelevant facts

It seems that both problems are, to a large extent, caused by the fact that the current variant of DIVERSUM does not allow for repeated selection of the same arc label in the given "zone". This property can be viewed as "extreme diversification" and actually seems to be too strong. As a consequence, when each arc label in the current zone is represented by some selected fact, the algorithm starts to look for facts that are in further zones, that would result in topic drift or presenting some irrelevant facts.

While PRECIS is totally diversity-oblivious, the notion of diversity in the DIVERSUM algorithm is extreme in the sense that it forbids *any repetition* of any arc label within a fixed arc-node distance from the summarised entity. We observed that this extremity sometimes causes another problem with the output. Namely, DIVERSUM cannot repeat arc label in given *zone* and thus is sometimes forced to explore triples that are distant by many hops from the summarised entity $q$. As such triples represent facts that do not directly concern $q$, we can call this problem as "topic drift" (known in information retrieval). Dealing with this phenomenon is out of scope of this article and is envisaged in our future work.

This problems can be addressed in an improved variant of the algorithm, by allowing for a mild arc label repetition in a given zone. In addition, such a "relaxation" of the algorithm may produce better results, since it would be possible to better represent the characteristic of a given entity by showing more facts that concern the main activity of this entity (e.g. more facts concerning "acted in" for a famous, active actor).

In future work it would be interesting to experiment with other types of entities (e.g., directors or movies) and on datasets from different domains.

The ALC diversity measure proposed in this article is a bit simplistic. It would be interesting to propose a more elaborated measure. Some initial observations are presented in Sydow (2011). Another direction is to work on the automated visualisation of the graphical entity summarisation results. Currently we use the AGNES algorithm presented in Sobczak et al. (2012). In an improved automated visualisation, the arc labels could be automatically grouped (e.g. facts concerning private life of an actor vs his professional life) based on some automatically collected statistics such as arc label co-incidence. It would be also very interesting to incorporate the notion of user interest profiles to the GES problem towards personalising the resulting summarisation.

The mentioned ideas are an object of our ongoing work and are envisaged to be studied in a separate publication.

We believe that this article is a good starting point to initiate a deeper discussion on the summarisation problem on semantic knowledge graphs, and on the concept of diversity in the domain of graphs in general.

**Acknowledgements** The work is supported by the N N516 481940 grant of the Polish Ministry of Science and Higher Education. The first author was supported by the DAAD Prize in 2010 for visiting MPII, Saarbruecken, where early part of this work was done. The authors would like to thank all the volunteers from the MPI for Informatics that participated in the expert-based experimental evaluation.

## Appendix: Selected user feedback collected in the experiments

16 comments of expert users that explicitly mentioned "diversity":

•*"Summary A has important and diverse relations summarising life and achievements of the actor", •"Facts with diverse relations in Summary A", •"There are only actedIn relationships in the summary B. Summary A is more diversified - it has a lot more information", •"more diversity, more information directly relevant to entity", •"diverse relations", •"diverse facts", •"diversity", •"diverse facts, not just actedIn", •"more diverse", •"diverse", •"diverse relationships in the left summary", •"more diverse", •"more diverse relations, but I dont like to many facts about related entities", •"more diverse", •"Too many actedIn facts in summary A. Summary B considers diverse relations", •"More diversity".*

9 comments of expert users that implicitly appreciated diversity-awareness:

•*"covers more topics, not just actedIn", •"summary B gives broader view of the entitys facts. Summary A is only about movies", •"summary A has a lot of interesting and different facts about the entity", •"more interesting facts: bornOnDate, livesIn", •"Left has mostly actedIn, too specific for summary", •"more types of facts", •"Gives important information about the actors life and 1-2 important movies", •"Summary A summarizes the life of entity Robert Mitchum and summary B summarizes his popular movies", •"more types of facts",*

Other positive comments on DIVERSUM referring to more diverse fact coverage

*Collected in the first crowd-sourcing experiment:*

•*"Summary B covers more aspects related to the entity" •"Summary A provides more information (for instance on the Date of birth) than Summary B." •"Because it gives only about films acted and only about a single film bride of frankenstein" •"in summary A describe details of his life events from birth to death" •"Summary A explains the life history of Denzel Washington. The graph briefly explains it." •"This algorithm shows the family concept.It is nice." •"Summary A has been explained with a detailed information of chevy Chase with his birth date and production year and* *with all his awards." •"In summary A useful facts are given. So its easy to have a quick summary. In B only the names of movies are given." •"It clearly explains the biography of Dharmendra." •"it contains also information about his life"*

*Collected in the second crowd-sourcing experiment:*

*•"Summary A contains more diversified information than Summary B which mainly focuses on the movies." •"It includes much more information regarding Denzels life while Summary A only includes movies he acted in." •"because summary a shows all the information including both the personal and acting career." •"Summary A has complete description of entity so it is preferred as an ad-hoc." •"Summary B covers almost all the facts needed to evaluate the entity from all aspects" •"Summary A just provided the brief outline about the films acted by the actor and nothing much about him." •"The Summary is more versatile and easily comprehendable" •"I think summary B gave a more overall abbreviated biography of his life" •"Summary B provides information that is of more interest to most. Majority of people know before hand what movies Tom Hanks was in. This was about him as more than just an actor" •"It has provided the required personal details in an apt manner." •"It is more understandable and exhaustive." • "this explains the details about Chevy Chase personal and official." •"Summary A is more clear and that there is a clear distinction between the actors professional and private life in this."*

*Negative comments on DIVERSUM collected in the crowd-sourcing experiments:*

*• "The summary A gives importance on Amitabhs son as well (like the 2 movies in which Abhishek Bachchan has acted in). However the summary B is very precise and gives importance only to Amitabh Bachchan.", • "Summary B has more information about the movie 'Die hard 2' than on the entity Bruce Willis", • "I get more better image in my mind about the actor with the summary of the films he has worked in or created from Summary B rather than his bio data from Summary A."*

## References

Abu-Jbara, A., & Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, HLT '11* (vol. 1, pp. 500–509). Stroudsburg, PA, USA. Association for Computational Linguistics.

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM international conference on web search and data mining, WSDM '09* (pp. 5–14). New York, NY, USA: ACM.

Alonso, O., Rose, D.E., Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum, 42*(2), 9–15.

Barzilay, R., & McKeown, K.R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics, 31*(3), 297–328.

Brabham, D.C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence, 14*, 75–90.

Carbonell, J., & 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*, (99 335–336). New York, NY, USA.

Carenini, G., Murray, G., Ng, R. (2011). *Methods for Mining and Summarizing Text Conversations*. Morgan & Claypool Publishers, 1st edn.

Chen, H., & 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* (pp. 429–436). New York, NY, USA: ACM.

Cheng, G., Ge, W., Qu, Y. (2011). Generating summaries for ontology search. In *Proceedings of the 20th international conference companion on World wide web, WWW '11* (pp. 27–28). New York, NY, USA: ACM.

Cheng, G., Tran, T., Qu, Y. (2011). Relin: relatedness and informativeness-based centrality for entity summarization. In *Proceedings of the 10th international conference on The semantic web - Volume Part I, ISWC'11* (pp. 114–129) Berlin, Heidelberg: Springer-Verlag.

Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: 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* (pp. 659–666). New York, NY, USA: ACM.

Dokulil, J., & Katreniaková, J. (2008). Visual exploration of rdf data. In *Proceedings of the 34th conference on current trends in theory and practice of computer science, SOFSEM'08* (pp. 672–683). Berlin, Heidelberg: Springer-Verlag.

Hovy, E.H. (2005). *Automated text summarization.* Oxford University Press, Inc., New York, NY, USA.

Elbassuoni, S., Ramanath, M., Schenkel, R., Sydow, M., Weikum, G. (2009). Language-model-based ranking for queries on rdf-graphs. In *Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09* (pp. 977–986) New York, NY, USA: ACM.

Goffman, W. (1964). A searching procedure for information retrieval. *Information Storage and Retrieval, 2*(2), 73–78.

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

Gupta, V., & Lehal, G. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258–268.

Ham, F., Schulz, H., Dimicco, J.M. (2009). Honeycomb: Visual analysis of large scale social networks. In *Proceedings of the 12th IFIP TC 13 international conference on human-computer interaction: part II, INTERACT '09* (pp. 429–442). Berlin, Heidelberg: Springer-Verlag.

Howe, J. (2008). *Crowdsourcing: why the power of the crowd is driving the future of business*. Crown Publishing Group, New York, NY, USA, 1 edn.

Kairam, S., MacLean, D., Savva, M., Heer, J. (2012). Graphprism: compact visualization of network structure. In *Proceedings of the international working conference on advanced visual interfaces, AVI '12* (pp. 49 8–505). New York, NY, USA: ACM.

Kittur, A., Chi, E.H., Suh, B. (2008). Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '08*. New York, NY, USA: ACM.

Kittur, A., Chi, E.H., Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '08* (pp. 453–456). New York, NY, USA: ACM.

Li, L., Zhou, K., Xue, G.-R., Zha, H., Yu, Y. (2009). Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web, WWW '09* (pp. 71–80). New York, NY, USA: ACM.

Li, P., Jiang, J., Wang, Y. (2010). Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual meeting of the association for computational linguistics, ACL '10* (pp. 640–649). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American chapter of the association for computational linguistics on human language technology, NAACL '03* (vol. 1, pp. 71–78). Stroudsburg, PA, USA: Association for Computational Linguistics.

Mani, I. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA, USA.

Penin, T., Wang, H., Tran, T., Yu, Y. (2008). Snippet generation for semantic web search engines. In *Proceedings of the 3rd Asian Semantic web conference on the semantic web, ASWC '08* (pp. 493–507). Berlin, Heidelberg: Springer-Verlag.

Ramanath, M., & Kumar, K.S. (2009). Xoom: a tool for zooming in and out of xml documents. In *Proceedings of the 12th International conference on extending database technology: advances in database technology, EDBT '09* (pp. 1112–1115). New York, NY, USA: ACM.

Ramanath, M., Kumar, K.S., Ifrim, G. (2009). Generating concise and readable summaries of xml documents. *CoRR*. abs/0910.2405

Robertson, S.E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294–304.

Sobczak, G., Pikuła, M., Sydow, M. (2012). Agnes: a novel algorithm for visualising diversified graphical entity summarisations on knowledge graphs. In *Foundations of intelligent systems, Proc. of 20th International symposium, ISMIS 2012, Macau, China, December 4–7, 2012* (vol. 7661, pp. 182–191). LNCS/Springer.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management, 43*(6), 1449–1481.

Sydow, M. (2011). Towards the foundations of diversity-aware node summarisation on knowledge graphs. In *Proceedings of "Diversity in Document Retrieval" workshop, european conference on information retrieval ECIR 2012, Dublin, Ireland*, (pp. 16–20).

Sydow, M., Pikuła, M., Schenkel, R. (2010). Diversum: Towards diversified summarisation of entities in knowledge graphs. In *Proceedings of data engineering workshops (ICDEW) at IEEE 26th ICDE Conference* (pp. 221–226). IEEE.

Sydow, M., Pikuła, M., Schenkel, R. (2011). To diversify or not to diversify entity summaries on rdf knowledge graphs? In *Proceedings of the 19th international conference on foundations of intelligent systems, ISMIS'11* (pp. 490–500). Berlin, Heidelberg: Springer-Verlag.

Sydow, M., Pikuła, M., Schenkel, R., Siemion, A. (2010). Entity summarisation with limited edge budget on knowledge graphs. In *Proceedings of the international multiconference on computer science and information technology* (pp. 513–516). IEEE.

Thalhammer, A., Knuth, M., Sack, H. (2012). Evaluating entity summarization using a game-based ground truth. In Cudr-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) *The semantic web ISWC 2012* (pp. 350–361). *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Thalhammer, A., Toma, I., Roa-Valverde, A.J., Fensel, D. (2012). Leveraging usage data for linked data movie entity summarization. *CoRR*. abs/1204.2718

Cormen, T.H., Leiserson, C,E., Rivest, R.L. (1990). "*Introduction to Algorithms*. MIT Press.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A. (2008). Efficient computation of diverse query results. In *Proceedings of the 2008 IEEE 24th international conference on data engineering, ICDE '08* (pp. 228–236). Washington, DC, USA: IEEE Computer Society.

Wackerly, D.D., Mendenhall, W., Scheaffer, R. (1990). *Mathematical Statistics with Applications*. PWS-KENT Publishing Company.

Waitelonis, J., & Sack, H. (2012). Towards exploratory video search using linked data. *Multimedia Tools and Application, 59*(2), 645–672.

Wan, X. (2009). Topic analysis for topic-focused multi-document summarization. In *Proceedings of the 18th ACM conference on information and knowledge management, CIKM '09* (pp. 1609–1612). New York, NY, USA: ACM.

Wan, X., Jia, H., Huang, S., Xiao, J. (2011). Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR '11* (pp. 735–744). New York, NY, USA: ACM.

Wan, X., Li, H., Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual meeting of the association for computational linguistics, ACL '10* (pp. 917–926). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wan, X., Li, H., Xiao, J. (2010). Eusum: extracting easy-to-understand english summaries for non-native readers. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10* (pp. 491–498). New York, NY, USA: ACM.

Wan, X., & Xiao, J. (2010). Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *ACM Transactions on Information and System, 28*(2), 8:1–8:34.

Wattenberg, M. (2006). Visual exploration of multivariate graphs. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '06* (pp. 811–819). New York, NY, USA: ACM.

Zhang, N., Tian, Y., Patel, J.M. (2010). Discovery-driven graph summarization. In *Proceedings of the 26th International conference on data engineering (ICDE)* (pp. 880–891).

Zhang, X., Cheng, G., Qu, Y. (2007). Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World wide web, WWW '07* (pp. 707–716). New York, NY, USA: ACM.