Semantic Entity Search Diversification

Tuukka Ruotsalo
Helsinki Institute for Information Technology HIIT
Aalto University, Espoo, Finland
Email: first.last@hiit.fi

Matias Frosterus
Department of Media Technology, School of Science
Aalto University, Espoo, Finland
Email: first.last@aalto.fi

***Abstract*—We present an approach to diversify entity search by utilizing semantics present and inferred from the initial entity search results. Our approach makes use of ontologies and independent component analysis of the entity descriptions to reveal direct and latent semantic connections between the entities present in the initial search results. The semantic connections are then used to sample a set of diverse entities. We empirically demonstrate the performance of our approach through retrieval experiments that use a real-world dataset composed from four entity databases. The results indicate that our approach significantly improves both diversity and effectiveness of entity search.**

I. Introduction

A significant proportion of Web searches are directed toward finding information about entities. Good examples of entities are databases managing information about products, points of interest, or artefacts – e.g., books, music, or images. Entity data on the Web are often accompanied with structured annotation that describes information such as authors, locations, or subject matter related to each entity. Therefore, entity search features specific challenges that set it apart from conventional Web search: 1) the entity collections are limited in size, 2) entities are often described with semantically similar but syntactically heterogeneous vocabulary, and 3) the user's information needs are often focused not on a single type of entity but on a set of related alternatives. As a consequence, standard information retrieval results in low recall and in entities that represent a limited set of subtopics, forcing the user to sift through a large number of irrelevant entities and miss many relevant ones [8].

For example, a user looking for information about entities when visiting a museum would limit the search to the collection specific to that particular museum and apply a variety of preferences characterising his or her tastes [18], [22], [16], [20]. For example, a user interested in natural sciences could specify her information needs by choosing to view entities associated with 'scientific instruments', 'astronomy', and 'natural sciences'. Conventional information retrieval, however, could lead the user to miss entities related to similar but not exactly matching entities, such as entities annotated with the concepts 'astronomers' (referring to scientists practising astronomy) and 'sundials' (a reference to a specific kind of scientific instrument related to astronomy).

Furthermore, conventional information retrieval methods could give a high ranking to entities that match several directly observable criteria – say, entities described in terms of the concepts 'astronomy' and 'astronomers'. This could cause a set of relevant entities with fewer matching criteria to have lower ranks and therefore lead to users easily disregarding these entities. In our example, entities described only via the concept 'sundial' could be ranked lower and ignored by the user even though still highly relevant for the information need. This phenomenon, called over specialisation, may reduce users' satisfaction with the retrieval results [14].

A promising solution to the over-specialisation problem is diversification of search results [1]. The basic premise of search result diversification is that the relevance of a set of documents or entities depends not only on the individual relevance of its members but also on how they relate to one another. Diversification methods are designed to increase the variety at the top of the result list by representing various aspects and interests expressed in a query. While diversification has recently become a popular research topic in the information retrieval community [1], entity search imposes additional challenges for diversification. The limited size, structured data, semantic heterogeneity, and over-specialisation of the initial results require advanced diversification methods.

We contribute a three-phase retrieval method to respond to these challenges. First, we employ ontology-based reasoning that allows semantic matching to increase recall of entity search. Second, we employ independent component analysis (ICA) to detect latent grouping of the entities in the result set. As opposed to similarity-based clustering, our approach makes use of higher-level statistics to detect latent variables shared by the entities and form groups of entities that are maximally independent from each other. This allows diversification based on latent structures and maximises independence, with measurement of diversity rather than similarity computed only on the basis of directly observable features. Third, we sample the most relevant entities from each independent component, thus providing a diversified view of results to the end user. We report results from laboratory experiments wherein a comprehensive real-life dataset was used in evaluation of our approach. We show that our approach increases the diversity of the entity search results and even improves the effectiveness of entity search. Because of the separate retrieval and post-diversification steps, our approach is also efficient and, on average, requires less than one second of computation for a resulting entity set.

978-0-7695-5119-7/19 $26.00 © 5119 IEEE
DOI 10.1109/ICSC.2013.16

IEEE computer society

II. Related Work

Most of the previous work on diversifying search results has focused on ambiguous and under-specified queries. The promise of diversification is to provide the user with results corresponding to diverse interpretations of the ambiguous query and removing duplicate interpretations [1]. The work has been conducted primarily in the scope of large digital library and Web data collections [1], [12], [5], [23], [2]. Some recent research has tackled the diversification problem with structured data sources [11].

The importance of result diversification was recognised already in early studies of information retrieval [6], but has been widely research only recently. but has become widely researched only recently. Most of the current work on result diversification makes at best only tacit use of the topics of the query or documents, and diversification occurs by way of similarity functions or conditional relevance distributions defined over documents [6], [26], [9] or through user feedback [17]. Our approach focuses on detecting independent components that are explained by latent variables. This allows us to detect relevant clusters of entities on the basis of a high-recall result set that consists of entities and their rich semantic descriptions.

Vee et. al. [25] study diversification in the context of structured data. Their key idea is to select a set of entities that are as diverse as possible according to a lexicographical ordering of features. The authors apply post-processing to retrieved results and offer two algorithms that directly take diversification into account. As the lexicographical preferences are known in advance and directly used in the problem formulation, their approach is a form of explicit diversification. Agrawal et. al. [1] point out that the notion of relevance is suppressed in the above-mentioned work, since the objective is to select entities that are distinct from one another, but that the method seems to work well for structured domains. In [10], the focus is on developing a framework of evaluation that takes into account both novelty and diversity and wherein questions and answers are treated as sets of information nuggets and relevance is a function of the nuggets contained in the questions and the answers. We apply a similar idea in our evaluation by sampling entities from clusters and evaluating the performance against the combined relevance assessments.

A promising way of coping with the over-specialisation problem involves grouping of the results [13]. A recent work by Carpinento [7] shows that diversification of top hits is more useful for quick coverage of distinct subtopics whereas clustering is better for full retrieval of individual subtopics, with a better balance in performance achieved through generation of multiple subsets of diverse search results. This is in line with our results and shows that when queries grow large (e.g., because of a semantic query or document expansion) and may contain many diverse entries, clustering performs well. Carpinento also concluded that there is little scope for improvement over the search engine baseline unless we are interested in strict full-subtopic retrieval. However, we found that when the search is recall-oriented, semantics and diversification can lead to significant improvements in both efficiency and the diversity of the entity search results. To our knowledge, our work is the first to examine simultaneously diversification and semantics in entity search, wherein content heterogeneity is predominant and structured and semantic descriptions of the entities are increasingly important.

III. Semantic Search and Diversification

We present a three-phase method for semantic entity search diversification that combines ontology-based reasoning and retrieval with post-retrieval diversification that employs independent component analysis. An initial ranking is performed through retrieval of the results using a vector space model (VSM) for information retrieval in a triple-space [19]. Post-retrieval diversification is then performed for the top-ranked entities in the initial result set. This ensures that the diversification is based on the features of the entities that are relevant for the initial information need get a high ranking and limits the number of entities and features thereof to be analysed in the post-retrieval step.

A. Data Representation and Indexing

We represent the entities by using the RDF(S) [4] language, which has become a de facto language for representation of entities on the Semantic Web. An example of an entity description from a collection of the Museo Galileo is shown in Figure 1. The entity description consists of literal values and values referring to ontologies that are used in reasoning. This allows reduction of sparseness between the representation of the queries and the representation of the entities. The ontologies are represented in RDF, and we index entities after a reasoning step. In other words, in the indexing phase we use a knowledge base that has gone through deductive reasoning in accordance with the RDF(S) semantics. The RDF(S) reasoning is performed by computing of the closure of an RDF input graph under the RDF(S) semantics with Horst partial closure that keeps the RDF(S) reasoning decidable [24].

In the indexing phase, the entities are represented as feature vectors in the Euclidean space. We use the RDF triple as an indexing unit and tokenise and stem the words in the literal fields by using Porter stemming.

The indexing of the entities and the triples in the deductive closures of their descriptions are represented as vectors describing the occurrences of each triple given it's property p. Now the vectors indexing the triples are created for each property as $V_p = (w_1, ..., w_k)$, where k is the number of all original and deduced triples for which the property p holds.

B. Initial Ranking

We use the $tf - idf$ weighting over the triples as described in [19]. Intuitively, the triples higher in the hierarchy accrue more occurrences through reasoning and their tf value will be higher, but their weight is reduced by the idf. This allows the popularity of a concept in the annotations to reflect the weighting of the triples [21]. For the most general match, where the property would be *rdf:Property*, the resulting vector space would be weighted using the $tf - idf$ weights for all concepts used for indexing, because it is the most general property. In this way, the heuristic for approximation of the triple's importance would be much lower than in the case wherein only the object that matches a specific property is used.

In the vector model, the triple vectors can be used for computation of the degree of similarity between each entity O stored in the system and the search criteria Q. The vector model evaluates the similarity between the vector representing an individual entity V_{O_j} and the search criteria V_Q. We reformulate the cosine similarity to take into account the set of vector spaces as suggested in [3], one for each property instead of using only one vector space for all of the triples:

$$sim(U,O) = \sum_{p=1}^{p} \frac{\sum_{i=1}^{k} v_{i,j_p} \cdot v_{i,q_p}}{\sqrt{\sum_{i=1}^{k} v_{i,j_p}^2} \cdot \sqrt{\sum_{i=1}^{k} v_{i,q_p}^2}}, \qquad (1)$$

where k is the total number of triples in each vector space p, j is an index for an entity, q is an index for a search criteria, i is an index for a triple, and p is an index for a property vector space.

C. Post-retrieval Diversification

The VSM gives us a one-dimensional ranking of the entities but is unable to detect whether the entities in the ranked list have latent connections that group them together and distinguish them from other entities. On the basis of latent grouping, results can then be diversified via sampling of entities from these groups. We apply ICA to obtain these latent connections [15]. We store the subset of matching columns and rows of the initial matrix used by the VSM into a new triple-entity matrix X.

We store only the weights w for matching triples and entities for which the cosine similarity is greater than 0: in other words, we store those triples that matched the query and elements within any of the entities.

The entities do not need to share any triples directly: ICA defines a generative model for the observed multivariate data that can reveal latent connections between the entities. In the model, the data variables are assumed to be linear mixtures of unknown latent variables, and the mixing is also unknown. The latent variables are assumed to be non Gaussian and mutually independent, and they, therefore, are called the independent components of the observed data [15]. These independent components can be found by the means of ICA. Formally, the ICA model [15] can be defined as:

$$X = AS, \qquad (2)$$

The rows in X represent the weights of triples that match the initial query and correspond to the columns that representing the entities. The matrix contains only the matching entities and triples appearing in the descriptions of the entities selected in line with the initial ranking determined using Formula 1. The unknown matrix S includes the original source signals, called independent components, and A is an unknown mixing matrix.

Because both the mixing matrix and the independent components are unknown, the model is required to estimate both A and S. In the case of a single row, given the model and the triples $x_1, \ldots, x_n$ of the random vector x, the task is to estimate both the mixing matrix A and the sources s. After estimation of matrix A, we can compute its inverse W, and obtain the independent components by $s = Wx$. We assign the entity to a component in accordance with the magnitude of the recovered source signal. These components are clusters that consist of the entities assigned to them.

Mixing matrix A can be estimated from looking for maximally non-Gaussian variables. This can be detected via adaptive calculation of the w vectors in W by estimation of the independence using, for example, mutual information or negentropy and setting up of a contrast function that either minimises mutual information or maximises negentropy [15].

We used the FastICA [15], an efficient ICA algorithm that has a fast convergence making it practical for our purpose that requires online performance. The hyperbolic tangent ($tanh$) was used as the contrast function to maximise negentropy. In preprocessing steps we first centre the data (force to zero-mean) and then whiten the data (force to uncorrelated components) and reduce the dimensionality by principal component analysis (PCA). Whitening ensures that all dimensions are treated equally by removing any correlations in the data. This is computationally demanding, but, because of the initially ranked matrix X, can be done in less than 0.5 seconds on a standard desktop machine. We use the eigenvalue filter value of 98% which means that after the PCA sorts the eigenvalues, the first highest eigenvalues, whose sum is above 98% of all of the eigenvalues are used. Additionally ICA must be provided with the maximum number of components desired, which we set to 10 after experimenting with the actual data. The contrast function value was set to $a = 1$.

The values for the contrast function and the eigenvalue filter were chosen over 1,000 runs, with different configurations, and choice of those values minimising the variance for the average entities assigned for the components. This was done because, in view of the purpose of the application, it was deemed desirable to produce components that are as even-sized as possible. Testing for the contrast function was done in 0.1-sized increments from 1.0 to 2.0 and for the eigenvalue filter in half-a-percentage-point-sized steps from 90 to 100. Using this procedure and these settings, we are able to keep the response time of the system to less than one second with a regular desktop machine. After the entities are assigned to the independent components, we rank the entities within the components on the basis of the original cosine similarity values. Finally, we rank the components in accordance with the highest cosine similarity occurring in the highest-ranking entity assigned in the component. This provides a total ordering for the results that can be used to represent the entities for the user. Total ordering is also used in the evaluation, wherein the quality of ranking of the entities is measured.

IV. Experiments

We conducted a set of experiments to determine the retrieval performance of the methods. We measured two elements: 1) whether the usage of independent component analysis maintained or improved the overall performance of entity search while at the same time introducing diversity in the results and 2) whether the performance was dependent on the higher recall enabled by the ontology-based reasoning. We used a 2 × 2 experimental design. The first condition was whether the ontology-based reasoning was used or not. The second condition involved whether independent component analysis was used or not.

A. Data

We used a dataset of slightly over 1,000 entities in the domain of cultural heritage. The entities have high quality structured annotations. The dataset consists of descriptions of entities such as museum items, including artwork, fine arts items, and scientific instruments, along with points of interest, such as visit spots, statues, and museums. The data were obtained from the Museo Galileo in Florence, Italy, and from Heritage Malta. While limited in size when compared to a traditional information retrieval benchmark collection, our collection is complete and contains all the entities exhibited in the museums and all cultural-heritage-related points of interest that exist for two cities. In this sense, it is comprehensive and realistic and can be used for evaluation of the methods for meeting real information needs. The entity annotations utilise the Dublin Core properties and required extensions for the cultural heritage domain, such as material, object type, and place of creation of the entity described. An example annotation of an entity describing a scientific instrument from the Museo Galileo is described in Figure 1.

The entities are indexed with RDF(S) versions of the Getty vocabularies[1]. The RDF(S) versions of the Getty vocabularies are large lightweight ontologies that are transformed to RDF(S) from the original vocabularies, wherein concepts are organised in subsumption hierarchies and have related term relations. The vocabularies consist of tens of thousands of concepts each. Geographical instances are structured in meronymic hierarchies that represent geographical inclusion. These were handled separately in the reasoning process through use of the part-of relations for geographical inclusion. Temporal data were handled similarly, through use of a handcrafted ontology that has concepts for each year, decade, century, and millennium organised in a hierarchy.

B. Queries and Relevance Assessments

The query set consists of 20 initial queries that were defined by domain experts at the same museums and cultural heritage institutions where the datasets were curated. Two types of queries were created: general entity queries and point-of-interest queries.

Figure 2 shows two examples of the queries, one for astronomers and the subject matter optics and the other for points of interest of the type villas, mansions, theatres, or palaces. Relevance assessments corresponding to the query set were provided for a set of 500 entities in both museums. Domain experts provided relevance assessments for the dataset by assessing each entity as either relevant or not relevant separately for all of the queries. The dataset and relevance assessments were carried out by the domain experts specifically for this study. This is a relatively large set of queries and relevance assessments for a one-off experiment: recall is analysed with full coverage by domain experts, meaning that all of the entities are manually inspected against all of the queries. Pooling or automatic pre-filtering was not used. This makes the relevance assessments highly reliable, avoids bias caused by automatic pre-filtering, and takes into account all possible semantic relevance – even non trivial connections judged relevant by the domain experts.

The domain experts were asked to create queries typical for the domain such that the queries would include also non-trivial queries considering the underlying collection. For example, a query including the concept 'seascapes' was judged relevant also for entities annotated with the concept 'landscape paintings' and for entities annotated with 'marinas', 'boats', 'harbours', etc. The judges were allowed to inspect the textual description in addition to the image of the entities when assessing relevance. A single set of query triples typically corresponded to a relatively intuitive subset of the entities. For example, the gold standard for a query related to the subject matter astronomers, astronomical photography, optical toys, and optical properties featured lenses, sundials, and telescopes but also biographies of astronomers. It is noteworthy that only a small proportion of the relevant entities in the gold standard are directly described using these triples but reasoning and latent analysis are required for matching of relevant entities. The gold standard for criteria of galvanometers, batteries, electrical engineering, and the name of famous physicist Leopoldo Nobili consisted of specific types of batteries, galvanometers, and other entities related to the concept of electrical engineering but also other entities relevant in relation to Nobili. It is significant that Nobili is related to many batteries and galvanometers but also to a large number of other entities. In this case, the traditional methods that are not able to cope with over-specialisation can fail completely, because they tend to return any entities related to batteries and electrical engineering designed by Nobili yet are not able to detect latent connections between entities that are typical for Nobili but not indexed as being a kind of galvanometer or battery. For investigation of the effect of diversification, we shuffled 40 new sets of queries in such a way that each resulting set of queries consisted of three of the original queries.

The rationale was to force various types of results in the result set and measure whether our diversification method would detect the subgroups automatically. This shuffling was done separately for site-specific collections and for the point-of-interest collection. We also combined the corresponding

[1] http://www.getty.edu/research/conducting_research/vocabularies/

```
<dc:identifier> <urn:imss:instrument:402015> .
<sm:physicalLocation> <http://www.imss.fi.it/> .
<dc:title> "Horizontal dial" .
<dc:subject> "Measuring time" .
<dc:description> "Sundial, complete with gnomon..." .
<dc:subject> <aat:300054534> . (Astronomy)
<sm:dateOfCreation> <sm:time_1501_1600> . (16th Century)
<sm:material> <aat:300010946> . (Gilt Brass)
<sm:objectType> <aat:300041614> . (Sundial)
<sm:placeOfCreation>  <tgn:7000084> (Germany)
<sm:processesAndTechniques> <aat:300053789> . (Gilding)
<dc:terms/isPartOf> "Medici collections" .
<rdf:type> <sm:Instrument> .
```

Fig. 1: An example of a description of an entity in the dataset used in the experiments. The subjects of the triples are all identifiers of the resource being described and, therefore, are omitted. The description is shortened.

```
<rdf:Property> <aat:300025789> . (astronomers)
<dc:subject> <aat:300134506> . (astronomical photography)
<dc:subject> <aat:300211119> . (optical toys)
<dc:subject> <aat:300056210> . (optical properties)
<rdf:type> <sm:Instrument> .

<rdf:type> <aat:300005517> . (villas)
<rdf:type> <aat:300071272> . (mansions)
<rdf:type> <aat:300007117> . (theaters)
<rdf:type> <aat:300005734> . (palaces)
```

Fig. 2: An example of two sets of queries defined by experts at the Museo Galileo. The namespace *dc* and *sm* refer to the Dublin Core and a custom extension of the Dublin Core properties for the cultural heritage domain, and *aat* to the Art and Architecture Thesaurus of the Getty Foundation. The subject of each RDF triple is omitted, because it is not bounded for these queries.

gold standards.

The VSM was set to return the 150 highest-ranking entities, and the diversification method returned the top 50 entities in such a way that it picked entities in equal numbers from each of the clusters on the basis of their original cosine similarity within the cluster. The rationale in this setting is that, if the diversification method is to perform as well as the VSM alone or better, it would need to have an effectiveness equal to or greater than that of the VSM. The VSM only ranked the entities in one dimension, while the ICA diversification raised the ranking of those entities that were originally ranked lower in the initially ranked list of the 150 entities. This is because they belonged to a component that ICA deemed relevant and thus were determined to belong to the new result set of 50 entities formed on the basis of the diversification process.

C. Evaluation Measures

Evaluation measures were selected to measure two types of performance: retrieval effectiveness and subtopic coverage. Precision and recall alone are vulnerable measures because often when precision increases, recall decreases and vice versa. Therefore, a single measure that can be used to estimate effectiveness that is balanced in terms of precision and recall can be useful. We are also interested in whether the ranking retained good accuracy even in the case in which entities were included for which the initial ranking yielded by the VSM was lower. A suitable measure for use to investigate the ranking along with the precision/recall trade-off is mean average precision (MAP). We allow self-organisation of the subtopics on the basis of latent variables relevant for the result set at hand. This is a departure from the approach of existing diversification measures, wherein the optimal diversified result ranking is pre-defined for each query in the gold standard. This is why we selected the Jaccard coefficient to measure subtopic coverage (the Jaccard coefficient measures similarity between sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets). When the value of the Jaccard coefficient is combined with MAP, the measures together yield the ranking and its similarity to the one returned by the retrieval method alone. The rationale is that diversity of the entities in the result sets is indicated by a low Jaccard coefficient value and similarity by a high value. At the same time, for the result set to be relevant, the MAP should be the same or better. A relatively low Jaccard

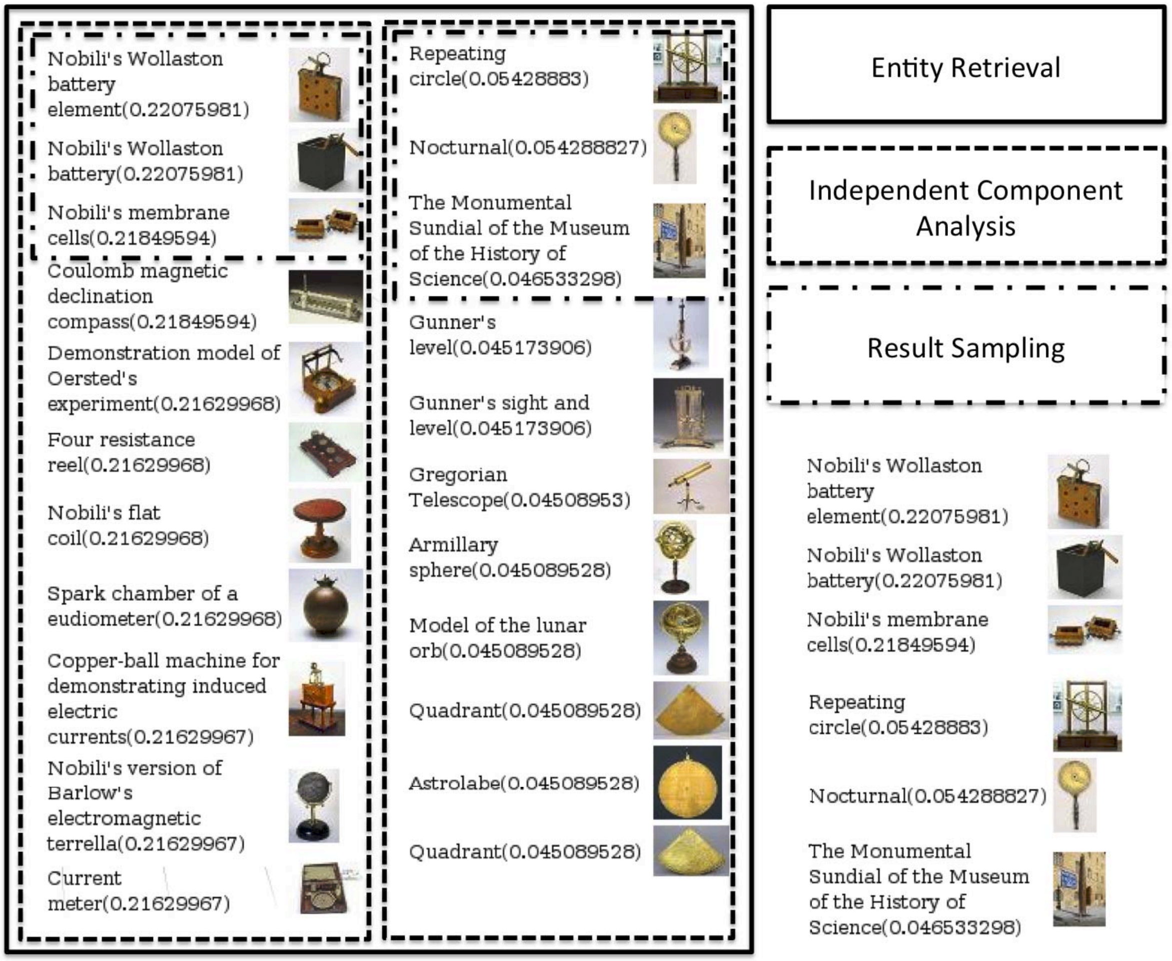

Nobili's Wollaston battery element(0.22075981)
Nobili's Wollaston battery(0.22075981)
Nobili's membrane cells(0.21849594)
Coulomb magnetic declination compass(0.21849594)
Demonstration model of Oersted's experiment(0.21629968)
Four resistance reel(0.21629968)
Nobili's flat coil(0.21629968)
Spark chamber of a eudiometer(0.21629968)
Copper-ball machine for demonstrating induced electric currents(0.21629967)
Nobili's version of Barlow's electromagnetic terrella(0.21629967)
Current meter(0.21629967)
Repeating circle(0.05428883)
Nocturnal(0.054288827)
The Monumental Sundial of the Museum of the History of Science(0.046533298)
Gunner's level(0.045173906)
Gunner's sight and level(0.045173906)
Gregorian Telescope(0.04508953)
Armillary sphere(0.045089528)
Model of the lunar orb(0.045089528)
Quadrant(0.045089528)
Astrolabe(0.045089528)
Quadrant(0.045089528)
Entity Retrieval
Independent Component Analysis
Result Sampling
Nobili's Wollaston battery element(0.22075981)
Nobili's Wollaston battery(0.22075981)
Nobili's membrane cells(0.21849594)
Repeating circle(0.05428883)
Nocturnal(0.054288827)
The Monumental Sundial of the Museum of the History of Science(0.046533298)

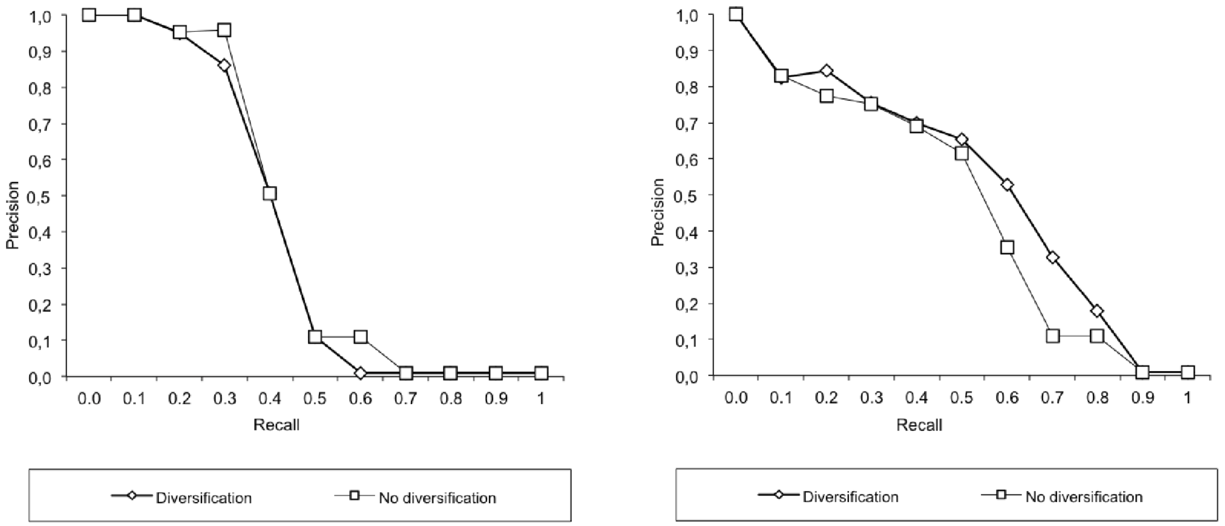
Fig. 3: The three-phase diversification process for an example query consisting of the concepts batteries, electrical engineering, Leopoldo Nobili, astronomy, and optics. First, semantic search is used to provide a high-recall result set of entities. Then, independent component analysis detects the latent components. The component on the left consists of entities related to batteries, electrical engineering, and Nobili, and the component on the right consists of entities related to optics and astronomy. The final result set (in the lower right-hand corner) is sampled by selection of entities with the greatest cosine similarities from each component.

coefficient coupled with a high MAP characterises a diversified and highly relevant result set.

D. Results

The experiments led to three main findings. First, the best performance was achieved with the combination of reasoning and diversification. The MAP for the retrieval with reasoned data was 0.48 and for the condition with reasoning and diversification was 0.53. Diversification improved the performance by five percentage points (11%). Second, the improvement was dependent on reasoning that enabled improved recall in the initial ranking phase. Diversification without reasoning was found to impair retrieval performance by two percentage points (4%). The MAP for the method with diversification was 0.41 and for the method without diversification 0.43 when reasoning was not used. Third, reasoning on its own improved the performance of the retrieval by five percentage points (12%). The results imply that our diversification technique is effective only when combined with initially high recall achieved using semantic reasoning. The results also demonstrate that diversification improves not only the diversity of the results but also the overall retrieval performance.

Figure 4a shows the precision-plotted-against-recall curve for the conditions wherein reasoning is not used, and the conditions with and without diversification are compared. The performance of the latent analysis, a condition wherein diversification is used, is lower than that for the condition that uses only retrieval. The recall of the retrieval method without reasoning is significantly lower than that in the case wherein reasoning was used, and precision drops rapidly at recall level 0.4. This implies that in the case wherein recall is low,



(a) No reasoning

(b) Reasoning

Fig. 4: Precision-recall curves for the measured conditions. Semantic reasoning increases the performance of the entity retrieval (a) compared to the condition in which reasoning is not used (b). Diversification improves the performance of the entity retrieval when reasoning is used (b), but has no effect when reasoning is not used (a). Precision is plotted on 11 recall levels. The values are computed as an average over the 40 shuffled queries.

the clustering hampers the overall performance of the entity retrieval. The results show that the tail of results achieved through semantic indexing allows effective entity retrieval.

Figure 4b presents the precision–recall curve for the conditions wherein reasoning is used, and the retrieval methods with and without diversification are compared. As expected, precision is high for both conditions at low recall levels. This can be explained by the fact that both methods are able to rank the highest-ranking entities correctly, because this is not affected by the post-retrieval diversification. Unlike in the case wherein reasoning was not used, the condition with diversification actually improves precision when recall increases. A possible explanation is that diversification allows omission of non-relevant results because the correct associations of the entities within the independent components convey the information about the relevance of the entities. In other words, the entities that are explained by the latent variables become selected into the final result list even though they were assigned a lower rank in the initial ranking. This implies that the diversification method is capable of improving retrieval performance when it is applied for initial results for which the recall is high and features are present for effective performance of the latent analysis.

The statistical significance of the differences between the conditions was verified via the following procedure: The data were not found to be normally distributed according to the Shapiro–Wilk test. Statistical significance was then ensured using the Friedman test, a non-parametric test based on ranks that is suitable for comparing more than two related samples. The statistical significance of the relations between conditions was then analysed via the paired Wilcoxon signed-rank test with post hoc test. The differences were found to be statistically significant ($p<0.001$).

We also manually examined the results for all 40 shuffled queries. An example of the diversification procedure is shown in Figure 4, wherein clear separation between entities related to electrical engineering and entities related to astronomy and optics can be observed. The Jaccard coefficient between the top 50 results directly returned by the semantic search and the top 50 results resulting from sampling from the independent components produced by the diversification method. The average Jaccard value was 0.31, meaning that the majority of the results in the case of post-retrieval clustering differ from those in the initial entity retrieval. This, in combination with the increased retrieval performance, indicates that our approach is effective in both diversifying results and improving retrieval performance. The average running time for the post-retrieval diversification process is less than a second, and the total response time with retrieval and sampling included came to around one second. This makes our approach suitable for real-world applications.

V. Conclusions

We studied the question of how semantic heterogeneity and over-specialisation of entity search can be reduced through diversification. This is a problem that most search engines targeted at search entities face, since the result sets can be described with heterogeneous vocabulary that may not match the queries users construct to represent their information needs. We took into consideration both the relevance of the documents and the diversity of search results, and we presented a three-phase diversification process wherein ontologies are used for semantic reasoning and independent component analysis for latent grouping of the results. An experiment with a comprehensive real-world dataset demonstrated that the combination of semantic reasoning and diversification consistently outperforms results produced by the entity retrieval method without deiversification. Our diversification technique also has other advantages over many of those proposed previously. We employ independent component analysis that groups the entities based on their latent connections on the basis of how independent they are, not how similar they are. An advantage of such a model is that it is able to detect latent variables that explain the data on a higher level than just similarity of the features of the entities. The true test of entity search engines is their ability to satisfy the complex and possibly contradictory information needs of their users. While our approach improves the retrieval performance and diversifies the results of entity search, we believe further work is necessary in designing user interfaces that are more reflective of the user needs in the highly varied entity search use cases.

References

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] Sumit Bhatia. Multidimensional search result diversification: diverse search results for diverse users. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1331–1332, New York, NY, USA, 2011. ACM.

[3] Roi Blanco, Peter Mika, and Sebastiano Vigna. Effective and efficient entity search in rdf data. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 83–97, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] D. Brickley and R. V. Guha. RDF vocabulary description language 1.0: RDF Schema W3C recommendation. Recommendation, WWW Consortium, Feb 2004.

[5] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. Efficient diversification of web search results. *Proc. VLDB Endow.*, 4:451–459, April 2011.

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

[7] Claudio Carpineto, Massimiliano Damico, and Giovanni Romano. Evaluating subtopic retrieval methods: Clustering versus diversification of search results. *Information Processing & Management*, 48(2):358 – 373, 2012.

[8] Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. A survey of web clustering engines. *ACM Comput. Surv.*, 41:17:1–17:38, July 2009.

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

[10] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, New York, NY, USA, 2008. ACM.

[11] Elena Demidova, Peter Fankhauser, Xuan Zhou, and Wolfgang Nejdl. Divq: diversification for keyword search over structured databases. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 331–338, New York, NY, USA, 2010. ACM.

[12] Zhicheng Dou, Sha Hu, Kun Chen, Ruihua Song, and Ji-Rong Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 475–484, New York, NY, USA, 2011. ACM.

[13] Marti A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49:59–61, April 2006.

[14] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[15] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626 –634, may 1999.

[16] Innar Liiv, Tanel Tammet, Tuukka Ruotsalo, and Alar Kuusik. Personalized context-aware recommendations in smartmuseum: Combining semantics with statistics. In *Proceedings of the The Third International Conference on Advances in Semantic Processing (SEMAPRO 2009)*, Sliema, Malta, October 2009. IEEE Computer Society. Sliema, Malta.

[17] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 784–791, New York, NY, USA, 2008. ACM.

[18] Tuukka Ruotsalo. *Methods and applications for ontology-based recommender systems.* Aalto University, School of Science and Technology, Ph.D. Dissertation, June 2010.

[19] Tuukka Ruotsalo. Domain specific data retrieval on the semantic web. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012*, Lecture Notes in Computer Science 7295, pages 422–436, Heraklion, Greece, 2012. Springer.

[20] Tuukka Ruotsalo, Krister Haav, Antony Stoyanov, Sylvain Roche, Elena Fani, Romina Deliai, Eetu Mäkelä, Tomi Kauppinen, and Eero Hyvönen. Smartmuseum: A mobile recommender system for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 20(0):50 – 67, 2013.

[21] Tuukka Ruotsalo and Eetu Mäkelä. A comparison of corpus-based and structural methods on approximation of semantic relatedness in ontologies. *International Journal On Semantic Web and Information Systems*, 5(4):39–56, 2009.

[22] Tuukka Ruotsalo, Eetu Mäkelä, Tomi Kauppinen, Eero Hyvönen, Krister Haav, Ville Rantala, Matias Frosterus, Nima Dokoohaki, and Mihhail Matskin. Smartmuseum: Personalized context-aware access to digital cultural heritage. In *Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, Trento, Italy, September 2009. Trento, Italy.

[23] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 881–890, New York, NY, USA, 2010. ACM.

[24] Herman J. ter Horst. Completeness, decidability and complexity of entailment for rdf schema and a semantic extension involving the owl vocabulary. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3), 2005.

[25] Erik Vee, Utkarsh Srivastava, Jayavel Shanmugasundaram, Prashant Bhat, and Sihem Amer-Yahia. Efficient computation of diverse query results. In Gustavo Alonso, José A. Blakeley, and Arbee L. P. Chen, editors, *ICDE*, pages 228–236. IEEE, 2008.

[26] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 504–511, New York, NY, USA, 2005. ACM.