Towards Faceted Search for Named Entity Queries

Sofia Stamou, Lefteris Kozanidis[1]
Computer Engineering and Informatics Department, Patras University, GREECE
{stamou, kozanid}@ceid.upatras.gr

Abstract. A considerable fraction of the web queries contain named entities. This, coupled with the fact that a proper name might refer to multiple entities, imposes the ever-increasing need that search engines handle efficiently named entity queries. In this paper, we present a technique that automatically identifies the distinct subject classes to which a named entity query might refer and selects a set of appropriate facets for denoting the query properties within every class. We also suggest a method that examines the distribution of the identified query facets within the contents of the query matching pages and groups search results according to their entity denotation types. Our preliminary study shows that our technique identifies useful facets for representing the named entity query properties in each of their referenced subject classes.

Keywords: faceted search, named entity queries, Wikipedia corpus.

1 Introduction

The key objective of all Information Retrieval systems is to help the users find the desired information about their search requests in an effortless yet successful manner. With the proliferation of the web content, search engines have become an indispensable tool in the information seeking process. Despite the popularity and the tremendous capacity that search engines have nowadays, there are still open issues concerning their ability to satisfy all user needs. This is essentially due to vocabulary mismatches between the indexed documents and the user issued queries that hinder the engines' ability in detecting the underlying correlation between documents and queries.

In this paper, we suggest a method for identifying the referred concepts of Named Entities (NE) in both user queries and query matching pages, in order to improve the engines' ability in handling named entity queries. What motivated our study is on the one hand the observation that a significant portion of popular web queries contain NEs [1] [2] and on the other the fact that different entities might be verbalized in user queries with the same name. As examples, consider the NE queries *Java* which might intend the retrieval of information about the programming language or the coffee, and *Apple* which might intend the retrieval of information about the company, the computer or the fruit. As the examples indicate, a NE query might refer to different entity

[1] The authors is financially supported by the PENED 03ED_413 research grant, co-financed by 25% from the Greek Ministry of Development-General Secretariat of Research and Technology and by75% from the E.U. European Social Fund.

types, but in the absence of any implicit *knowledge* about the different reference classes of named entities, the engine would always retrieve the same results; usually a mixed list of pages about the distinct query denotations. Consequently, users would have to go over a long list of results and access their contents in order to satisfy their search intent.

To overcome such difficulties, researchers have proposed a number of techniques for personalizing search results according to specific user interests. Personalization, although it might work well for some users or queries, nevertheless it entails practical limitations in a real deployment as it pre-requisites processing large volumes of web transaction logs in order to model the user search preferences. Another possible direction towards helping users find the desired information is to classify web pages into faceted hierarchies and present the query matching results grouped together according to the corresponding query senses [7]. Although faceted search was primarily addressed in the context of cataloguing and library systems, nowadays it is widely employed by e-commerce applications (e.g. amazon.com, shopping.com) and recently it has attracted the interest of the web search community. This surge of interest is basically because facets allow a document to exist simultaneously under different categories, each representing distinct document properties, and enable the user access the categorized documents in multiple ways.

Bearing in mind the power and success of existing faceted search approaches, we introduce the use of facets for describing the different entity types to which NE queries might refer and for detecting within the contents of every query matching page the specific subject that query entities denote. To enable that, we start by defining a set of facets that are good candidates for representing the different types of NE queries. In this respect, we rely on the Wikipedia corpus[2] and a number of heuristics for annotating every NE query with one or more suitable facets. Then, we examine the NE query results in order to estimate the distribution of the query facets within the contents of every query matching page. Based on the underlying association between the facets of the queries and their returned pages, we propose grouping search results according to the query entity types that their contents represent. Grouped search results accompanied by their derived faceted terms, when displayed to the users can help them find the information sough in an effective yet efficient manner. Our preliminary study shows that our technique delivers useful facets for representing the subject classes that NE queries might represent. Thus, we claim that our approach could be fruitfully explored towards improving the users' search experience when querying the web for named entities.

The remainder of the paper is organized as follows. We begin our discussion by reviewing related work. In Section 3, we present our approach towards the automatic identification of useful facet terms for representing the NE query classes. In Section 4, we propose a faceted search approach that groups query results according to the representation of the query facets in their contents. In Section 5, we report the results of our experimental evaluation and we conclude the paper in Section 6.

[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download

2 Related Work

Many researchers have studied the problem of named entity recognition and disambiguation. To address the problem, researchers have proposed numerous ways for exploring entity-local features [3] [4], or complex lexico-syntactic and morphological data [5] [6] in order to detect NEs within text documents. Recently, [13] [7] suggested the use of encyclopedic knowledge for resolving the different classes to which a NE might refer. In their work, [7] utilize the Wikipedia corpus in order to derive a dictionary of named entities and then based on the detected entities' contextual elements and position in the Wikipedia taxonomy, they determine a set of candidate concepts that every entity denotes. They disambiguate NEs based on the degree of correlation between the entities' contextual elements and their conceptual categories. In a different approach, [8] studied the enrichment of Wikipedia pages with named entities. In this respect, they designed a method that extracts a number of features (both contextual and page-based) from the Wikipedia pages and uses them as training examples for a classification module. Upon training, the classifier assigns every Wikipedia page with an appropriate tag (from a pre-defined list) that is reflective of the named entities contained in the page. In a similar task, [9] employed a graph-based approach and experimented with categorizing named entities in the Japanese version of Wikipedia. Recently, [17] proposed a model for representing web pages as sets of named entities that are generally informative of the following subject types: *person*, *location*, *organization* and *time*. Based on the inter-relations between these subject types, the authors determine which named entities suffice for representing the content of web pages.

Our work also relates to some recent studies on faceted search and facet terms extraction. More specifically, in [10] and [11] the authors introduce a method for identifying useful facets for browsing textual databases. Their method relies on WordNet [12] hypernyms, Google search results and the Wikipedia pages for defining a set of broad terms with which to expand keyword terms extracted from the database contents. In [14] the authors propose the faceted query logs analysis and present a method for classifying web queries into the following faceted categories: *ambiguous*, *authority*, *temporally-* and *spatially-sensitive* requests. In [15] a faceted search personalization approach is discussed, emphasizing on how to customize the search interface according to user ratings. In [16] a data-driven technique, called Dynamic Category Sets (DCS), is introduced that discovers sets of values across multiple facets that best match the query and enables the user disambiguate the latter via a search-by-category clarification dialog. In a recent study, [18] introduce a dynamic faceted search system that automatically discovers a small set of valued facets that are deemed interesting to a user and enable the latter understand the important patterns in the query results.

Although our study touches upon issues that have been previously addressed, it is different from existing works in the following: we introduce a novel facet extraction technique that is specifically tailored for representing the classes of NE queries. Our method uses a small amount of NE contextual data from which it extracts useful facets. To estimate the usefulness of the identified faceted terms, we propose a metric that estimates how valuable a facet is for representing the named entity properties within a particular subject class. We also suggest the exploitation of the identified NE query facets while looking for query relevant data. In this respect, we suggest a method that relies on the distribution of both the query keywords and the query facets

within the query matching pages in order to group search results according to their named entity denotations. In the following section, we discuss the details of our facet selection approach in order to represent the subject denotations of NE queries.

3 Faceted Representation of Named Entity Queries

In this section, we discuss the details of our work towards associating named entity queries with a set of useful facets for denoting the properties of their refereed classes. There are two main challenges associated with our goal: how to identify all possible subject classes to which a NE query might refer and how to select useful terms for verbalizing the NE query properties in each of the identified classes. For the first challenge, we explore the Wikipedia corpus and we apply a number of heuristics for determining a set of features that are informative of the NE subject classes (Section 3.1). For the second challenge, we introduce a facet extraction algorithm that explores the entities' contextual elements in order to derive useful terms for describing the NE properties in each of the referenced classes (Section 3.2). Before delving into the details of our method, let's describe the process we adopt for identifying NE queries.

For any query *q*, we firstly wish to determine whether *q* is a named entity. To be able to judge that, we rely on the Wikipedia pages in the titles of which we look for the query keywords. Upon their detection, we download the contents of the respective pages and we follow the steps suggested by [7] in order to assess whether a given term corresponds to a NE or not. These steps summarize to (quoting from [7]): (i) *If q is a multiword query, check the capitalization of all content terms in q. If all contents words of q appear always capitalized in their corresponding pages then q is a NE*. (ii) *If q is a single-term query and contains at least two capital letters, then q is a NE*. (iii) C*ount the occurrences of q terms in the text of the page, in positions other than at the beginning of the sentences. If at least 75% of these occurrences are capitalized, then q is a NE*.

The combination of the above steps in the work of [7] resulted to the extraction of nearly half a million named entities form the Wikipedia corpus. In our study, we apply the same process in order to automatically identity whether a search request is a named entity query or not. Note that this NE detection heuristic applies to languages that do not capitalize common nouns, e.g. English. Based on the identified NE queries, we designed a method that automatically derives the different subject classes to which a named entity might refer. The details of our method are discussed next.

3.1 Deriving the Named Entity Query Classes

The first step towards discriminating between the different subject classes to which a named entity query might refer is to examine whether the identified NE has a unique type of reference or not. In particular, we need to investigate if a NE always refers to a single class of subjects (e.g. humans, locations, etc.) or not. For our investigation, we start by looking at the presence of clarification sentences in the Wikipedia articles that discuss the given NE. Clarification sentences always appear at the beginning of a Wikipedia article right after the title section and are of three generic types, stating:

a) ***This article is about*** [`clarification term/phrase`]
b) ***For other uses, see*** (`link to a disambiguation page`)
c) ***For other*** [`clarification term/phrase`] ***with the same name, see*** (`link to a disambiguation page`)

In some cases, a clarification sentence might be a combination of the above types; usually of types (a) and (b). Based on clarification sentences, we can distinguish two main groups of NEs: those which are clarified in Wikipedia (i.e. their corresponding articles contain clarification sentences) and those which are not clarified.

For named entities with clarification sentences, we examine the type of their clarifications and we further sub-group them according to the following criterion: entities whose clarification sentences are all of type (c) listed above, are deemed as entities with a single reference class, where the latter is expressed via the clarification term(s). We refer to such entities as **single faceted NE** and we further process them to verbalize their facets as we will present in the following section. Named entities with clarification sentences of types (a), (b) or a combination of the above-listed types are intuitively considered to be expressive of multiple subject classes and we refer to them as **multi-faceted NE** (one clarification per subject class). Note that under our approach a disambiguation sentence[3] is generally deemed as a clarification one, assuming that Wikipedia contributors edited such sentences for clarifying the usage of a given entity. Therefore, a NE is deemed to be referring to as many classes as the number of the sentences in which the entity's name is clarified. In the following section, we present how to select useful faceted terms for denoting the different classes to which a NE might refer.

Now, let's go back to the group of named entities that lack clarification sentences in their corresponding Wikipedia articles. To identify which of these NE are single faceted and which are multi-faceted, we extract the definition sentences from their respective Wikipedia articles. A definition sentence is the first sentence in the body of the article that contains the named entity and a wordform of the verb *to be* followed by one or more common terms and/or disambiguation entities[4]. Relying on the extracted definition sentences, we compute their feature vectors within a sliding window of ten words surrounding the entity reference. For our computations we employ a bigram model that records for every word (i.e. feature) in a definition sentence its relative position around the entity and the number of times the word appears in that position. Then, based on the sentences' overlapping features in the derived vectors, we organize them into clusters of shared elements. For example, a significant portion of the Wikipedia definition sentences about humans, exhibit the feature (`born`) right after the entity's name. Under our criterion, all definition sentences containing the above feature are grouped together. Following the above steps, we organize every Wikipedia definition sentence about a NE into one or more clusters, depending on the number of their shared features.

Based on the number of clusters to which a named entity's definition sentence is assigned, we make the following assumption: if the definition sentence of a NE be-

[3] A disambiguation page in Wikipedia is a page that contains the term "disambiguation" in its title and contains several possible disambiguations of a term [7].

[4] According to [8], a disambiguation Wikipedia entity is a hyperlinked term that points to another Wikipedia page in which the meaning of the term is resolved.

longs to a single cluster, then the NE is single faceted. For example, a NE whose definition sentence is organized under a single cluster, which groups sentences sharing the feature t = *holiday* (e.g. Christmas is an annual *holiday*) is deemed as expressive of a single class of subjects. Conversely, if the definition sentence of a NE belongs to multiple clusters, then the NE is multi-faceted and refers to as many classes as the number of clusters to which the definition sentence of the NE has been assigned. For example, a NE whose definition sentence is grouped into (say) two clusters, one of which groups sentences that share feature t_i = *comic book* (e.g. *Superman is a fictional comic book superhero …. cultural icon*) and the other groups sentences that share feature t_j = *cultural icon* (e.g. *Superman is a fictional comic book superhero widely considered to be one of the most recognized of such characters and an American cultural icon*) is deemed as being expressive of two object classes, e.g. book and cultural icon. Based on the process described above, we can discriminate between named entities that always refer to a single class of subjects and those that might represent different subject classes in their contextual denotations. Following on from that, we need to define a set faceted terms that are useful for representing the NE properties in each of the identified subject classes. In this respect, we have designed a facet selection algorithm; the details of which are discussed next.

3.2 Selecting Faceted Terms to Represent the Named Entity Query Classes

In selecting a set of useful facets for describing the different types of named entity references, we rely on the contextual elements in the NEs' clarification and definition sentences, in which we look for faceted terms that represent the NE's class properties. The reason for relying on the contents of the Wikipedia articles for deriving the NEs facets instead of exploring the Wikipedia categories for those NEs, summarize to the following. First, not all Wikipedia articles about NEs have been associated with a category label. Moreover, most of the articles about NEs that are topically annotated have been assigned to numerous categories, whose class denotations and interrelations are not clearly distinguishable. Lastly, facets unlike topics represent the class properties of a concept (i.e. NE) rather than its semantic orientation. Therefore, in our work we decided to select the facets that describe the NEs class properties from the terms collectively selected by the Wikipedia editors for clarifying and/or defining NE.

Therefore, we address the challenge of identifying useful facets for representing the NE properties as a term extraction problem for which we need a sound model that accurately detects valuable faceted terms within the NEs' surrounding context. The greatest difficulty associated with implementing such a model is how to acutely detect which of the NEs' contextual features are good facets for describing the entity's properties. By good facets, we denote the terms that help us discriminate between the different named entity references while at the same time they are not overly specific, they are not redundant across different entity types and they are useful in a web search setting. To build a model that complies with the above criteria, we have designed a facet extraction algorithm, the basic steps of which are illustrated in Fig.1. In brief, our algorithm takes as input a set of named entities E for which we would like to define useful facets, and a set of clarification and definition sentences S that have been extracted from the Wikipedia pages that correspond to each of the above named

entities. Based on the above data and following the below-listed steps, our algorithm delivers for every class of named entities a sorted list of facets *F* that are useful for describing the named entity properties within their reference classes.

```
Input: Wikipedia clarification and definition sentences S, set of Named Entities E
Output: useful facet terms (F)
For each s in S do
  Extract all content terms T
      For each t in T do
        Compute R(t)
        Associate every t with corresponding entity e_i
      end
end
For each entity e_i in E collect all content terms associated with e_i, T_i (e_i)
    /*Compute terminological similarity*/
    For each e_i, e_j with terms Ti, Tj do
        S_term (e_i,e_j)
        If S_term (e_i,e_j) > 0.5
            Group e_i, e_j together
        end
end
For each e_i in class c
    Take all content terms t associated with e_i
    For each t in c do
        Compute R(t,c)
        Compute Usefulness (t)
    end
end
return top k-terms in Facet (F), ranked by Usefulness (t)
```

Fig. 1. Identifying useful NE faceted terms in the contextual elements of Wikipedia articles.

The first step of our approach extracts all content terms from the clarification and definition sentences (S) that have been collected for each of the named entities. For content terms extraction, we apply tokenization and Part-of-Speech tagging to all the sentences contained in *S* and we retain only the sentence nouns and proper nouns, based on the observation of [19] that terms of the above grammatical categories communicate most of the thematic properties of the text in which they appear. Then, for every extracted content term *t*, we estimate the fraction of named entities that contain *t* in their definition and clarification sentences. For our estimations, we firstly associate every term extracted from a sentence *s* to the respective entity *e* that is clarified or defined in the contents of *s*. Such *content term-named entity* associations can be easily derived from the associations between the named entities and the Wikipedia sentences about the entities that contain *t*. Then, we compute for every term *t* a value R(t) that indicates the "representation ratio" of *t* with respect to some entity, given by:

$$R(t) = \left| E(t) \right| / \left| E \right| \quad \textbf{(1)}$$

Where |E(t)| is the count of all named entities that contain *t* in their clarification and definition sentences and |E| is the count of all named entities identified. The value of R indicates how representative is a term for describing named entity properties,

assuming that as the value of R increases so does the term's probability of being a good facet for named entities. At the end of step 1, we associate every named entity with a set of content terms extracted from the Wikipedia sentences that describe the entity properties. Each of the extracted terms is also associated with an overall representation value, i.e. R score.

In the second step we measure the amount of overlapping content terms between named entity pairs, so as to be able to determine for every pair of named entities e_i, e_j whether they refer to the same class of subjects or not. In this respect, we rely on the content terms associated with every named entity (in step 1) and we compute the similarity (S_{term}) between named entity pairs as follows:

$$S_{term}(e_i, e_j) = \frac{2 \bullet |\text{common terms}|}{|\text{terms about } e_i| + |\text{terms about } e_j|} \quad (2)$$

The similarity between the named entities (e_i, e_j) content terms, takes values between 0 and 1; with zero indicating that the two entities have no content terms in common in their clarification and definition sentences and one indicating that all the terms in the named entity sentences are common.

Based on the above formula, we can derive for any pair of named entities the degree to which their contextual elements overlap. We then determine whether any two named entities refer to the same class of subjects based on the following criterion:

$$e_i, e_j = \begin{cases} \text{same class} & \text{if } S_{term}(e_i, e_j) > 0.5 \\ \text{different class} & \text{otherwise} \end{cases} \quad (3)$$

This way, we group named entities into subject classes according to their content terms' similarity values. Note that under our criterion, a named entity e_k might be grouped under several classes depending on the number of named entities with which e_k shares a significant amount of content terms. Having grouped NEs according to their contextual overlapping features, the next step is to determine a set of terms for representing the subject denotations of the named entities that are grouped together. More specifically, we need to identify among the content terms that have been determined for each of the NEs that refer to the same class, the ones that make the most useful and descriptive facets for representing the NEs' properties within that class.

In the last step, our algorithm determines useful faceted terms for each of the named entity classes as follows. Given a set of named entities that refer to the same subject class (i.e. they are grouped together) it collects all their content terms and starts by estimating a new *R* value for every term within a class, as follows:

$$R(t,c) = |E(t,c)| / |E(c)| \quad (4)$$

Where |E(t, c)| is the count of all named entities that refer to class *c* and which contain *t* in their clarification and definition sentences and |E(c)| is the count of all the named entities identified for class *c*. This new R(t,c) value indicates the "representation ratio" of *t* with respect to some class *c* so that terms with increased R(t,c) are better candidates for representing the class properties and consequently the NEs that refer to that class. Having computed the degree to which a content term *t* represents the properties of the entities that refer to a given class (i.e. R(t,c) value) and consider-

ing that we have already estimated (in step 1) the degree to which *t* makes a good term for representing NE properties, we easily derive the usefulness of *t* as a facet, as:

$$\text{Usefulness}\ (t) = R(t) \bullet R(t, c) \quad (5)$$

Based on the above formula, we estimate the usefulness of a term *t* in serving as a facet for some named entity that refers to a class *c* as the product of the term's representation ratio for named entity properties and the term's representation ratio for subject class properties. At the end of this process, we retain the top-*k* terms in each of the subject classes as the faceted terms that are useful in denoting the properties of the named entities that refer to that class.

4 Faceted Search for Named Entity Queries

So far we have presented our method towards identifying both the number of classes to which a named entity query refers and a set of useful facets for representing the query properties within every identified class. One last issue that our study addresses is how to be able to identify the query subject denotations within the contents of the retrieved pages. In this respect, we suggest an approach that examines the distribution of the query facets in the contents of the query retrieved pages and groups search results according to their query denotation types. In particular, given a NE query and a set of facets identified for each of the referring query classes our method employs a simple string matching approach and looks for the query faceted terms in the contents of the search results. In case a page contains some of the query facets, our module investigates whether these pertain to one or more subject classes. If all detected query facets represent a single class, then the faceted term of the highest usefulness value for that class is selected as the facet that represents the NE query denotation in the pages contents. The selected facet serves as a tag that is displayed next to the page in the search engine results. Conversely, if the query facets that a page contains represent multiple classes, then our module examines the position of the identified facets with respect to the query keywords in the page's content and selects the facet that is closest to the query terms as the one that represents the subject denotation of the NE query in the page. Again, the selected facet is used as a tag that indicates the query class representation in the page's content. Finally, in case a page does not contain any of the query facets, it receives no tag and as such it cannot inform the user about the NE query denotations. However, in the latter case, we suggest that query facets are displayed together with search results and enable the user click on the facet term that best suits her query intention. The selected facet is then appended to the query keywords and the refined query is re-submitted to the engine. Following the annotation of the query retrieved pages with an appropriate faceted term from the ones that have been identified for a NE query, we suggest grouping search results by subject denotations (i.e. faceted terms) and display them to the users accompanied by their identified tags. Based on this enriched list of search results, the users can make informed clicking decisions and satisfy their information needs faster.

5 Experimental Evaluation

In this section, we present the experimental evaluation of our facet extraction algorithm and we discuss obtained results. Due to the absence of a standard benchmark for evaluating the usefulness of the automatically selected NE facets, we carried out a human study, in which we measured the accuracy of the facets identified by our algorithm. For our experiment, we relied on 7,000 randomly selected named entities and their corresponding Wikipedia articles that served as our experimental data. Given that some of the NEs are discussed in more than one Wikipedia articles, the total set of the NE pages that we examined in our study is 19,350.

Following the method presented in Section 3.1, we processed the above data in order to determine for every NE the number of referenced subject classes. From all the NEs in our dataset, 410 had a single reference class and the remaining 6,590 had multiple reference classes (between 2 and 7). Then, we relied on the Wikipedia clarification and definition sentences that we extracted from the articles discussing each of our experimental NEs and we supplied them as input to our facet selection algorithm. The latter, following the steps discussed in Section 3.2, grouped the NEs into subject classes and for every class it identified a set of useful facets for denoting the NE properties within that class. In total our algorithm computed 12 subject classes for grouping our experimental NEs and for every class it selected k (k=10) faceted terms (i.e. a total set of 120 facets). The most useful facets in each of the identified classes are: *Person*, *Institute*, *Holiday*, *Country*, *Corporation*, *Novel*, *Disease*, *Newspaper*, *Science*, *Film, Band* and *War.*

To asses the performance of our algorithm in selecting useful NE facets, we carried out a human study in which we evaluated the accuracy of the facets that our algorithm selected for denoting the NE reference classes. To conduct our study, we recruited 45 volunteers from our school to whom we presented the list of NEs and their corresponding 19,350 Wikipedia pages and asked them for every NE to read the respective pages and indicate a set of terms that was in their opinion useful for representing the subject denotation of the NE within every page. We asked our participants to select up to 10 terms for every page referring to a NE and we indicated that terms may or may not appear in the contents of that page. Each of the 19,350 Wikipedia pages was examined by five participants and we considered a manually defined faceted term to be valid if at least three of the participants selected the same term to represent the NE reference in the page. In total, our participants identified 269 distinct facets of which 204 were selected by at least three different users. We then relied on these 204 jointly selected distinct facets and we compared them to the 120 facets that our algorithm selected for the same set of named entities and reference pages.

For our comparisons, we relied on the OSim measure [20] and computed for every NE the degree of overlapping facets between those selected by our participants and those delivered by our algorithm, Formally, the overlap between two lists of facets (each of size k) for a given NE is determined as:

$$\mathrm{OSim}\,(F_{manual}, F_{system}) = \left| T_{manual} \cap T_{system} \right| / k \qquad \textbf{(6)}$$

Where T_{namual} is the set of NE faceted terms that our subjects indicated, T_{system} is the set of NE faceted terms that our algorithm selected and k is the number of facets

considered, which in our case k=10 since both our participants and our algorithm delivered up to 10 facets for every NE. Table 1 lists obtained results. Due to space constraints, we report the number of our experimental named entities that exhibit similar degrees of overlapping facets across the different OSim levels. As the Table shows, our algorithm managed to identify for 95% of the NEs (i.e. for 6,650 out of the 7,000) at least one facet that was identical to a human selected one. A close look at the obtained results reveals that for NEs with highly overlapping facets, our participants picked faceted terms from the contents of the named entity pages. This demonstrates our algorithm's ability in identifying within the NE contextual elements the terms that are highly representative of the named entity denotations.

Although OSim is a useful measure for deriving the level of agreement between the human and the system selected facets, nevertheless it is marginally informative about how people perceive the usefulness of the automatically selected NE facets. To be able to judge that, we asked our participants to examine the facets selected by our algorithm and indicate which of these (if any) were in their opinion useful for representing the corresponding NE references in the contents of their pages. Human judgments on the usefulness of the automatically selected facets were binary (i.e. useful, non-useful) and as in the case of OSim, every facet was examined by five participants and we considered a facet to be useful if at least three of the users marked the facet as such. We list obtained results in Table 2.

Table 1. NEs distribution at overlapping similarity levels between the user-defined and the system-selected facets.

#of Named Entities	% of overlapping facets
350	0
785	0.1-0.2
806	0.2-0.3
1,965	0.3-0.4
1,622	0.4-0.5
778	0.5-0.6
424	0.6-0.7
170	0.7-0.8
80	0.8-0.9
20	0.9-1

Table 2. Statistics on the human judgments about the usefulness of the automatically selected facets.

# of algorithm selected facets	120
# of useful facets	86
# of non-useful facets	34
algorithm's success rate	71.7%

As Table 2 demonstrates most of the facets that our algorithm selects are deemed as useful by our participants, yielding an overall 71.7% algorithm success rate in automatically selecting useful named entities facets.

Overall, the results of our human study indicate that users consider the facets that our algorithm selected for representing the NE reference classes as useful. Given our algorithm's potential, we believe that it can be fruitfully employed in a web search setting in order to represent the named entity properties within both the user queries and the query retrieved pages. By using facets to represent both the query and the pages' subjects, we believe that users will be able to locate pages of interest faster and conveniently. In this respect, and considering that NE facets can be computed offline, we deem that our method can operate on-the-fly and be readily explored in web search applications.

6 Concluding Remarks

We presented a method that automatically identifies terms in the contextual elements of NE queries that are representative of the query subject denotations. We have also introduced a metric that estimates the usefulness of every identified facet for a given query reference class. Our experimental evaluation indicates that the facets selected by our algorithm are useful for denoting NE properties and as such we believe that they can be employed towards improving web searches about NEs. Currently, we are testing our algorithm's efficiency on a different corpus in order to validate how much the dataset used for the extraction of NE facets influences the quality of the obtained results.

References

1. Li, X., Liu, B., Yu, Ph: Mining Community Structure of Named Entities from Web Pages and Blogs. In AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.
2. Pasca, M.: Weakly-Supervised Discovery of Named Entities Using Web Search Queries. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, 2007.
3. Cucerzan, S., Yarowsky, D.: Language Independent NER Using a Unified Model if Internal and Contextual Features. In Proceedings of CoNLL Conference, pp. 171-174. 2002.
4. Klein, D., Smarr, J., Nguyen, H., Manning, C.D.: Named Entity Recognition with Character Level Models. In Proceedings of the CoNLL Conference, 2003.
5. Fleischman, M., Hovy, E.: Fine Grained Classification of Named Entities. In Proceedings of the COLING Conference, pp. 267-273. 2002.
6. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named Entity Recognition through Classifier Combination. In Proceedings of the CoNLL Conference, pp. 168-171. 2003.
7. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In Proceedings of the EACL Conference, pp. 9-16. 2006.
8. Dakka, W., Cucerzan, S.: Augmenting Wikipedia with Named Entity Tags. In Proceedings of the 3rd Intl. Joint Conference on Natural Language Processing. 2008.
9. Watanabe, Y., Asahara, M., Matsumoto, Y.: A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields. In Proceedings of the EMNLP-CoNLL Conference, pp. 649-657. 2007.
10. Dakka, W., Dayal, R., Ipeirotis, P.: Automatic Discovery of Useful Facet Terms. In Proceedings of the ACM SIGIR Workshop on Faceted Search. 2006.
11. Dakka, W., Ipeirotis, P.: Automatic Extraction of Useful Facet Hierarchies form Text Databases. In Proceedings of the ICDE Conference, 2008.
12. Fellbaum, Ch.: WordNet: An Electronic Lexical Database. MIT Press, 1998.
13 .Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proceedings of the EMNLP Conference. 2007.
14. Nguyen, B.V., Kan, M-Y.: Functional Faceted Web Query Analysis. In the WWW 2007.
15. Koren, J., Zhang, Y., Liu, X.: Personalized Interactive Faceted Search. In the WWW 2008.
16. Tunkelang, D.: Dynamic Category Sets: An Approach for Faceted Search. In Proceedings of the ACM SIGIR Workshop on Faceted Search. 2006.
17. Di, N., Yao, C., Duan, M., Zhu, J.J-H., Li, X.: Representing a Web page as Sets of Named Entities of Multiple Types – A Model and Some Preliminary Applications. In Proceedings of the World Wide Web Conference (poster session), pp. 1099-110, 2008.
18. Dash, D., Rao, J., Megoddo, N., Ailamaki, A., Lohman, G.: Dynamic Faceted Search for Discovery-Driven Analysis. In Proceedings of the 17th Intl. ACM CIKM Conference. 2008.
19. Gliozzo, A., Strapparava, C., Dagan, I.: Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. In Computer Speech and Language, 18(3):275-299, 2004.
20. Haveliwala, T.: Topic Sensitive PageRank. In Proc. of the 11th WWW Conference, 2002.