Searching for Individual Entities: a Query Analysis[*]

Barbara Bazzanella
University of Trento
DISCOF
Trento, Italy
barbara.bazzannella@unitn.it

Heiko Stoermer
Fondazione Bruno Kessler
DKM Unit
Trento, Italy
hstoermer@gmail.com

Paolo Bouquet
University of Trento
DISI
Trento, Italy
bouquet@disi.unitn.it

Abstract

Searching for information about individual entities such as persons, locations, events, is an important activity in Internet search today, and is in its core a very semantic-oriented task. Several ways for accessing such information exist, but for locating entity-specific information, search engines are the most commonly used approach. In this context, keyword queries are the primary means of retrieving information about a specific entity. We believe that an important first step of performing such a task is to understand what type of entity the user is looking for. We call this process Entity Type Disambiguation. In this paper we present a Naive Bayesian Model for entity type disambiguation that explores our assumption that an entity type can be inferred from the attributes a user specifies in a search query. The model has been applied to queries provided by a large sample of participants in an experiment performing an entity search task. The beneficial impact of this approach for the development of new search systems is discussed.

1 Introduction

In the transition from a "document web" to a "semantic web", one of the most significant changes in paradigm is the shift away from documents as the central element of information retrieval, towards something closer to the actual information need of the user. Neglecting navigational and transactional queries [4] in the context of our work, we follow the assumption that informational queries can be satisfied by identifying which individual entity a user is looking for information *about*. Studies have shown that user behaviour is often characterized by defining a certain context in which the desired information is most likely to be found [15], and from our perspective, an individual entity can be such a context.

Our work is thus concentrating on the question how to understand such a context, i.e. how to determine from a set of keywords whether there is a part that describes an individual entity, and which kind of entity it describes, in order to limit the search to information about this precise entity. We argue that identifying the entity target (and its type) in a query would help us to understand search intent better, and therefore provide better search. This approach is particularly meaningful for searching in Semantic Web content, where "aboutness" is a central aspect of information modelling. Knowing about what we want to know something can help us limit the search space significantly and improve the quality of search results. A first step in this direction was a study that we have performed in 2008, with the aim of asking people how they actually *describe* entities. Results of this study were published in [2], and provided us with a first hypothesis on the most important set of features commonly employed by users. This study has lead to the implementation of a novel algorithm for entity linkage [12], specially tailored for use-cases of the Semantic Web, as well as provided the background for the core data model in the Entity Name System [1].

The topic of this paper is a new study that has been conducted to gain insights into the same questions from a different perspective (user queries instead of descriptions), to see whether our initial findings can be confirmed, and to explore whether it is possible to identify an "entity part" in a keyword query. The analysis of the outcomes of this study are significant for the Semantic Web community for several reasons. First, we confirm many of the findings of [2], which represents a useful contribution to the ontological modelling of entity types, because we provide an extensive list of the most common features used to describe entities; these can directly be re-used to create or even to evaluate an ontological model. Secondly, as mentioned, search on the Semantic Web is more and more going in the direction of question answering, and understanding which (type of) entity we are

[*] This work is partially supported by the by the FP7 EU Large-scale Integrating Project **OKKAM – Enabling a Web of Entities** (contract no. ICT-215032). For more details, visit http://www.okkam.org.

978-1-4244-8099-9/10/$26.00 ©2010 IEEE

talking about can be important in this process. Finally, our findings can help us disambiguate an entity-related query; to a human, the term “Washington” in the two queries “George Washington” and “Washington USA” is clearly referring to different objects. We are hoping to give a contribution to the construction of new algorithms that also make this possible in a machine.

2 An Entity Search Experiment

Queries for specific entities represent a variation of the expressed information need that has been studied in many IR contexts [11, 14]. A query for a specific entity can be considered like a way to translate a human information need into a small number of attributes that the user considers relevant to identify the entity. Therefore, the analysis of real user queries should provide valuable insights into which kinds of attributes humans actually consider relevant to identify different types of entities during the search process.

As a first step towards a better understanding of this aspect of the query formulation process, we performed an experimental study. This study is part of an ongoing research to better understand how people represent and identify entities. In the previous experiment mentioned above, we adopted a bottom-up approach to investigate how people extensively describe individual entities belonging to a small set of entity types. The selection of these entity types was driven by a set of ontological requirements. In this current study we focus on the same collection of entity types, partly also to confirm our initial findings. Therefore, we refer the reader to the technical report [2] accompanying the first study for more theoretical details.

The goal of this study was to investigate the process that leads users to organize and represent their information needs using simple queries, limiting the analysis to queries that look for specific type of entities (person, organization, event, artifact and location). More specifically, this study explores two main issues: 1) to investigate which attributes are considered more relevant by people to identify specific types of entities in a query formulation task; 2) to identify significant patterns of attributes that reproduce recurrent strategies in entity searching.

2.1 Methodology

To answer our research questions we conducted an online experiment with a significant amount of users (301 participants with average age of 31.40 years and standard deviation of 9 years). The advantage of the online modality is twofold. First, the target of our research is a user population that has experience with Web-based information retrieval systems and the Internet provides a “natural” environment to reach this target. Second, the Internet experiment allowed us to access to a more diverse pool of participants (demographically and culturally).

The experiment consists of ten query formulation tasks. Participants are presented with an entity type (e.g., person) and they are asked to imagine any individual entity of their choosing belonging to this type (e.g., Barack Obama). Once the individual entity is chosen, participants are asked to formulate a query with the intent to find the homepage or an official Web site dedicated to the entity considered. In our example a plausible query may be <Barack Obama president USA>. Every participant is asked to perform ten such tasks, submitting their queries through a mimicked search engine interface . Five tasks present entity types at a very high level of abstraction. We call these types *high-level entity types* (person, organization, event, artifact and location). All the participants were tested on all the high-level tasks. The other five tasks correspond to more specific entity types (*low-level entity types*), selected from a predefined set of possible subtypes for each high-level type. Every participant performed only one low-level task for each high-level entity type. The task order was randomized between subjects. In the table 1 we report the complete list of high-level and corresponding low-level types.

Person	Organization	Event	Artifact	Location
politician	company	conference	product	tourist location
manager	association	meeting	artwork	city
professor	university	exhibition	building	shop
sports person	government	show	book	hotel
actor	agency	accident	article of clothing	restaurant
		sports event		

Table 1. Entity types and subtypes

2.2 A Naive Bayes Model of Attribute Relevance

The first goal of our research is to identify which kinds of attributes humans consider relevant to identify different types of entities during the search process. To answer this question we suggest to adopt a Naive Bayes Model of attribute relevance. This choice is motivated by two main reasons. The first is that quantifying the level of relevance of a feature for a category is a well-known approach in cognitive studies on human categorization [9]. Moreover, the Bayesian model of attribute relevance corresponds to one of the measures proposed in cognitive psychology [8] (cue validity) to quantify the relevance of a feature for general categories. A second reason is that the formalization of Bayesian statistics provides a middle ground where cognitive models and probabilistic models developed in other research fields (statistics, machine learning, and artificial intelligence) can find the opportunity for communication and integration.

In order to clarify the terms of our approach, we first introduce the basic tokens of the Naive Bayesian Model

(NBM). We can represent a query Q as a set of unknown terms $T = (t_1, t_2, ..., t_n)$, each of which can be a single word or a combination of words. We assume that each term t specifies the value of an attribute a. Assume that $A = (a_1, a_2, ..., a_n)$ is a set of attribute types. We map every term t into one appropriate type in A. Finally, suppose that $E = (e_1, e_2, ..., e_m)$ is a small number of entity types.

Our goal is to quantify the relevance of an attribute type a for a given entity type e_i. In the NBM framework this corresponds to compute the posterior probability $p(e_i|a)$:

$$p(e_i|a) = \frac{p(e_i) * p(a|e_i)}{\sum_{i=1}^{m} p(e_i) * p(a|e_i)} \tag{1}$$

The NBM is a probabilistic model based on the assumption of strong independence between attributes that means that the presence (or absence) of a particular attribute is unrelated to the presence (or absence) of any other attribute. Under this assumption, we can extend the model to the case of multiple attributes $\mathbf{a} = (a_1, a_2, ..., a_s)$ as defined in Eq. 2.

$$p(e_i|\mathbf{a}) = \frac{p(e_i) * p(\mathbf{a}|e_i)}{\sum_{i=1}^{m} p(e_i) * p(\mathbf{a}|e_i)} \qquad \text{where}$$

$$p(\mathbf{a}|e_i) = \prod_{j=1}^{s} p(a_j|e_i) \tag{2}$$

In this way, we can express the combined relevance of two or more types of attributes for a given entity type (e.g., the combined relevance of "name" and "surname" for "person") and detect the most likely type of the entity which the user is looking for.

Preprocessing Before applying the Bayesian Model to our data we performed two steps of preprocessing (see table 2 for examples). The first step, (*syntactic preprocessing*), involved extracting the terms from the queries. (Terms could be a single word or a combination of words). In this phase we also cleaned the dataset from unusual queries such as blank queries (empty), strings with only punctuation marks or senseless queries. Once the terms have been extracted from the queries, they were mapped into the attribute type set A. This mapping corresponded to the second step of preprocessing (*semantic preprocessing*). The first step was conducted in a semiautomatic way (i.e., the deletion of empty queries and a rough tokenization by segmenting the text at each space were performed automatically but the assignment of words to terms was performed manually), whereas the semantic preprocessing was performed entirely manually.

Query	**Syntactic Preproc.**	**Semantic Preproc.**
Q_1 =SWAP 2008 Rome	t_1=SWAP t_2=2008 t_3=Rome	$t_1 \Rightarrow$ event name $t_2 \Rightarrow$ date:year $t_3 \Rightarrow$ city
Q_2= McCain Republican	t_1=McCain t_2=Republican	$t_1 \Rightarrow$ surname $t_2 \Rightarrow$ political party

Table 2. Two-step Preprocessing

3 Results

In our experiment we collected an amount of 4017 queries. The average query length was 2.04 terms (mode=2 and median=2), which is in line with the results reported in literature (see for example [6]). Over 35% contained only one term and less than 3% of the queries contained five or more terms. Almost none of the queries utilized Boolean operators (over 99%). In only ten queries the operator AND was used, whereas the use of other operators was inexistent. The analysis of the word frequency showed a very limited usage of articles, prepositions, and conjunctions as demonstrated by the distribution of the high-frequency words. The only word without content that appeared in the first 30 most frequently used words was the preposition "of".

3.1 Bayesian Relevance of Attribute Types

In table 3 we report the results of applying the Bayesian Model described in Eq. 1 for the five high-level entity types addressed in our experiments. For each entity type we list the attributes with the highest relevance.

Entity Type (e)	**Attribute type** (a)	$p(e\|a)$
Person	first name	0.85
	surname	0.84
	occupation	0.89
	middle name	0.69
	pseudonym	0.33
	area of interest/activity	0.21
Organization	organization type	0.88
	organization name	0.73
	area of interest/activity	0.54
	url extention	0.29
Event	event name	0.96
	event type	0.95
	date:month	0.83
	date:year	0.81
	date:day	0.75
Artifact	artifact type	0.98
	features	0.90
	model name	0.89
	artifact name	0.86
	historical period/epoch	0.56
	nationality	0.50
Location	location type	0.84
	location name	0.65

Table 3. Bayesian Relevance: entity types

These attributes satisfy two requirements: they are frequently used by subjects to formulate their queries about a specific entity type and at the same time they are rarely used to search for other entity types. Therefore, if we are able to identify the relevant attributes, we can use this information to infer the entity type of the target entity. To test our prediction, we used 125 queries (25 for each of the five entity types) which were randomly extracted from the original query sample and we tested our model[1]. The results in terms of precision, recall and F-measure[2] are summarized in Table 4. The overall performance of our approach is satisfactory and promising, even though the evaluation measures show some interesting differences between the categories. The best result was obtained for the entity type “Event”, whereas we obtained the weakest result for the type“Artifact”. These results can be explained considering an important aspect of the identification strategies people use in making queries. Some entity types are more frequently used to specify the context in which a target entity is placed. For example, a product can be identified with reference to the company name, an artwork can be searched with reference to its author. Therefore, the presence of some attribute types, like “organization name” or “person name” in our example, may weaken the performance of our disambiguation method because they are not unique for specific entity types. On the contrary, since events are rarely used as contexts for other entity types, the presence of event attributes strongly indicates that the query is about an event.

Measures	Person	Organization	Event	Artifact	Location
Precision	0.72	0.87	1	1	0.85
Recall	1	0.91	0.91	0.66	0.96
F-measure	0.84	0.89	0.95	0.80	0.90

Overall Precision	**Overall Recall**	**Overall F-measure**
0.86	0.89	0.88

Table 4. Test-set Evaluation

Once the general type has been identified, a second step is discriminating between entities inside the same high level entity type. This problem can be formulated as follows. Which are the attributes that are relevant for certain entity subtypes belonging to the same high-level type? In order to answer this question, we performed the same analysis, restricting the domain to the low-level entity types. In table 5 we report an example of this second level of analysis (for space reasons we report only an extract of the results).

From an overall analysis of the results it stands out that for the majority of entity types “name” is the most relevant attribute used by people to identify the target of their request. This result confirms the centrality of proper names within the referential expressions (see for example [7]). However not all entities can be identified by means of a name. For example, pieces of clothing, accidents, or governments are entity types identified preferentially by means of other attributes. A particular case is represented by the entity type “product”. Our analysis shows that the majority of products are identified by the “model name” and not by the proper name of a specific entity (a type-token issue not uncommon in the context of products [10]).

Entity Type (e)	**Attribute type** (a)	$p(e\|a)$
Tourist location	location name	0.74
	location type	0.28
	organization name	0.18
City	administrative role	0.68
	building name	0.68
	state name	0.48
	municipality	0.48
	country name	0.46
	city name	0.30
Shop	shop name	0.91
	product type	0.90
	brand	0.85
	shop type	0.79
	address:street	0.33
Hotel	hotel name	0.93
	hotel type	0.61
	number of stars	0.48
	price range	0.42
Restaurant	restaurant name	0.92
	type of cousine	0.90
	restaurant type	0.61
	services	0.47
	neighbourhood	0.43

Table 5. Bayesian Relevance: Location

This result reveals another important aspect of the identification process: only a subset of entities are prototypically *namable* entities (e.g. person). Since users need also to identify non-namable things in their queries, the problem of Entity Type Disambiguation can not be entirely solved by the detection of the named entity in a query and the classification of it into predefined classes (an example of this approach can be found in [5]). Given a query like “guitar Jimi Hendrix 1967”, the named entities are “Jimi Hendrix” and “1967”, but the target entity of the query is an artifact (the guitar). The example shows that the simple classification of the named entities can be uneffective to detect the type of the target entity of the query and supports the idea that the disambiguation process can be performed combining different kinds of attribute.

The typology of the entity, for example, is a recurrent attribute specified by users. For example, for the entity type “organization” people frequently reported the “organization type” such as non profit, voluntary and so on. Artifacts are frequently identified non only by means of “the artifact type” but also by means of a variety of features, like qualitative attributes (e.g. “color” or “material”), “size” or “functions”. The identification by author is largely used for some subtypes of artifact such as “artwork” and “book”. Events are identified by means of both temporal and spatial attributes, whereas locations are preferentially described in terms of spatial attributes. Other attributes are distinctive of

[1] The 125 queries constituting our test set were not part of the sample which was used to calculate the Bayesian Relevance Measures reported in table 3.

[2] *Precision* is defined as the number of queries correctly assigned to the entity type divided by the total number of queries assigned to that type; *recall* is defined as the number of queries correctly assigned to the entity type divided by the total number of queries which should have been assigned to it; *F-measure* is the harmonic mean of precision and recall.

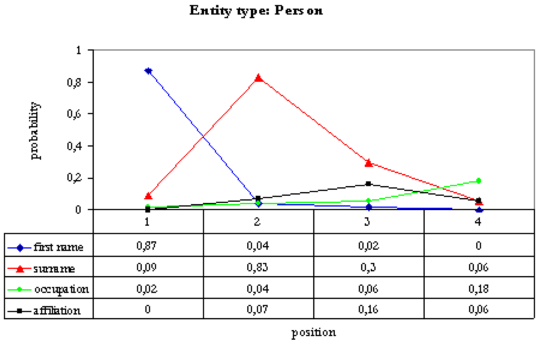

Figure 1. Distribution of attribute position

one subtype such as "publisher" and "editor" for "book" or "type of cuisine" for "restaurant".

3.2 Distribution Trends

The second research question of our study was about the distribution of attributes inside the queries. The idea was to highlight possible trends of attributes that recur during the formulation process. To this purpose we conducted two different kinds of analysis. First, we studied the distribution of attributes in terms of position. Second, we adopted a measure (Jaccard coefficient) to estimate the co-occurrence of the most relevant attributes.

3.2.1 Distribution of Attribute Position

As already mentioned, a query formulation process is highly selective: people pick a very small set of terms to express their information need in a suitable way to submit to an IR system. Despite the brevity of queries, a relevant aspect of querying behavior is about the strategies used by users to organize this information.

The first aspect that we investigated is about the position of attribute types within the query. The question can be formulated as follows: is there a preferential order followed by subjects when they organize the attributes within the query so that an attribute type is more likely reported in a specific position in the query? For example, is the name of the entity target always the first attribute specified? In this case the position of the attribute becomes extremely informative to understand the entity search process and should be included in an integrated model of attribute relevance.

The results of our experiment give support to this hypothesis. A significant example is reported in figure 1 that shows the probability distribution of the attribute types for the entity type Person. We note that "first name" is the attribute with the highest probability in first position, whereas "surname" is the preferred attribute in second position.

3.2.2 Co-occurrence of Attribute Types

The Bayesian model presented in section 3.1 by definition assumes the stochastic independence between attributes. Due to this (simplified) assumption, any possible correlation between attributes is ignored. However dependencies among attributes could emerge, providing further information to be exploited by automatic techniques of entity disambiguation. For example, the presence of an attribute type would be used to predict the type of another attribute, improving the disambiguation process. In order to produce a preliminary evaluation of this aspect we analyzed the co-occurrence of attribute types in queries.

As a measure of co-occurrence, we used the Jaccard coefficient [13] that captures the degree of co-occurrence of two objects (in our analysis two attributes).

Assume we are to measure the co-occurrence of two attributes a_1 and a_2. The number of queries containing both attributes is denoted by $|A_1 \cap A_2|$. Therein, A_1 denotes a query set that includes a_1 and A_2 denotes a query set that includes a_2. Then, the co-occurrence of a_1 and a_2, denoted by $J(a_1, a_2)$, is approximated by the Jaccard coefficient defined in Eq. 3.

$$J(a_1, a_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} = \frac{|A_1 \cap A_2|}{|A_1| + |A_2| - |A_1 \cap A_2|} \quad (3)$$

In Table 6 we report an example of the co-occurrence values for the entity type Person. Every cell in the tables contains the co-occurrence value for the correspondent pair of attributes. We note that the co-occurrence of the "first name-surname" pair is very high, indicating that these two attributes are more likely used in combination to identify persons in queries.

	first name	surname	mid-name	occupation	area-int.
first name	1	0.89	0.04	0.10	0.03
surname	0.89	1	0.04	0.11	0.04
mid-name	0.04	0.04	1	0.02	0
occupation	0.10	0.11	0.02	1	0.05
area-int.	0.03	0.04	0	0.05	1

Table 6. Attribute Co-occurence: Person

3.3 An example of Entity Type Disambiguation

A sketch of a possible application of the results of our study can be the analysis of the queries Q_1:< George Washington> and Q_2:<Washington USA>. In both we have the same term <Washington> to be disambiguated. But since in Q_1 the term <Washington> is preceded by the term <George> that is likely to be a "first name"we can use the measures of relevance, position and co-occurrence to infer that the type of attribute to assign to that term is more likely "surname" than "city name". From the results of our analysis we know that the attributes "name" and "surname" are the most relevant for the entity "Person" and they have an high value of co-occurence. Moreover, from the distribution analysis (see fig.1) "name" is more likely to be the first term specified in the query, whereas "surname" has the highest probability for the second position.

But how can we disambiguate that "Washington" in Q2 has a different meaning from "Washington" in Q1? From our data we know that when people search for a city, "city name" is one of the most relevant attribute (see table 5), as well as the most likely attribute in the first position. Moreover, a location is frequently identified by means of another location. From these evidences, we conclude that "Washington" in Q2 is more likely to be "city name" and the entity type of the query is "Location".

4 Conclusion

In this paper we have presented a cognitive experiment performed by a significant amount of participants, with the aim of investigating how people search for individual entities. This study is motivated out of specific needs that arise from ongoing work on an entity name system for the Semantic Web [3], where identifying a specific entity in a large entity repository based on a user query is one of the key problems, but also semantic search of information about individual entities is addressed.

The conclusions we draw from the data we collected in the experiment are several. First of all, we were able to confirm earlier findings from a different type of experiment which was performed to find out how people *describe* entities [2]. The combination of the results of both studies provide a community of ontology creators with a good background on how to model a certain set of entity types.

Second, we were able to extract certain patterns in the way people search for entities. One type of pattern is the typical position of a certain type of feature in a sequence of search terms (e.g. the fact that usually "first name" appears before "surname"). This result is relevant to tasks that have the objective of mapping keywords from a search query to a formal representation of a query that can be run against a system managing structured data (such as querying and RDF/OWL KB with SPARQL). Another type of pattern is the co-occurence of features in a query. We have established a catalog of typical co-occurences for a selection of entity types. These co-occurences can play a significant role in the disambiguation of queries, e.g. for solving the problem of "George Washington" vs. "Washington USA" mentioned in the introduction.

It is important to note that these findings are not limited to use cases of the Semantic Web, but we believe that especially the interdisciplinary view of an area between cognitive sciences, information retrieval and semantic systems on the Web can be a helpful contribution to future developments in the Semantic Web.

References

[1] B. Bazzanella, T. Palpanas, and H. Stoermer. Towards a general entity representation model. In *Proceedings of IRI 2009, the 10th IEEE Internationational Conference on Information Reuse and Integration, August 10-12, 2009, Las Vegas, USA*. IEEE Computer Society, August 2009. to appear.

[2] B. Bazzanella, H. Stoermer, and P. Bouquet. Top Level Categories and Attributes for Entity Representation. Technical Report 1, University of Trento, Scienze della Cognizione e della Formazione, September 2008. http://eprints.biblio.unitn.it/archive/00001467/.

[3] P. Bouquet, H. Stoermer, C. Niederee, and A. Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25, pages 554–561. IEEE Computer Society, August 2008.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval SIGIR 09 (2009)*, 2009.

[6] B. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52:235–246, 2001.

[7] S. Kripke. *Naming and Necessity*. Oxford, Basil Blackwell, 1980.

[8] E. Rosh and C. Mervis. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.

[9] G. Sartori and L. Lombardi. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16:439–452, 2004.

[10] P. Simons and C. Dement. *Aspects of the Mereology of Artifacts*, pages 255–276. Kluwer, Boston, 1996.

[11] R. S.Taylor. Process of asking questions. *American Documentation*, 13:391–396, 1962.

[12] H. Stoermer and P. Bouquet. A Novel Approach for Entity Linkage. In *Proceedings of IRI 2009, the 10th IEEE Internationational Conference on Information Reuse and Integration, August 10-12, 2009, Las Vegas, USA*. IEEE Computer Society, August 2009. to appear.

[13] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[14] R. S. Taylor. Question-negotiation and information-seeking in libraries. *College and Research Libraries*, 29:178–194, 1968.

[15] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422, New York, NY, USA, 2004. ACM.