Technical Report
2009
Searching for Individual Entities: a Query Analysis [*]

Barbara Bazzanella[2], Heiko Stoermer[1], and Paolo Bouquet[1]

[1] University of Trento,
Dept. of Information Science and Engineering,
Trento, Italy
{bouquet, stoermer}@dit.unitn.it
[2] University of Trento
Dept. of Cognitive Science
Trento, Italy
b.bazzanella@email.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 [16], and from our perspective, and 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. 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 [13], 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 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 [12, 15]. 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. 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 were 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. 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). The other five tasks corresponded to more specific entity types (*low-level entity types*), selected from a predefined set of possible subtypes for each high-level type. The task order was randomized between subjects. In the table 2.1 we report the complete list of high-level and corresponding low-level types.

Participants submitted their queries through a mimicked search engine interface as shown in Fig. 1.

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

Fig. 1. Screenshot of the Search Interface used in the experiment.

To receive more complete queries from the participants, in half of the tasks they received (after having submitted the query using the proper field) a request to refine the query. The other tasks were "one-shot", i.e. the refinement was not requested.

At the end of the experiment, the participants were asked to fill in an optional questionnaire about their background information (gender, age, country, Web and search engine experience).

The participants were invited to take part in our experiment via email[3]. 301 people participated in the study (165 male, 101 female, the others did not provide personal information). The average age of the participants was 31.40 years (SD=9). The experiment was performed in two versions, English (133 participants) and Italian (168 participants). In table 2.1 we report the distribution of the number of subjects that specified their native country (262 out of 301).

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

---

[3] To spread the participation request we submitted our post to mailing lists such as DBWorld or SIG-IRList.

Country	N	Country	N
Italy	155	Greece	5
Germany	18	Argentina	3
Spain	9	Austria	3
Ireland	8	Romania	3
Brazil	7	Belgium	3
Slovenia	7	Cambodia	3
Russia	6	Bulgaria	2
India	5	China	2
Others	21	USA	2
	N_{tot}	262 (out of 301 participants)	

Table 2. Geographical provenance of participants.

relevance. This choice is motivated by two main reasons. The first is that quantifying the level of relevance of a feature for a category[4] is a well-known approach in cognitive studies on human categorization [9, 10]. 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 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.

---

[4] The general idea is that among all the semantic features of a concept, those with high relevance are useful in discriminating the concept from other concepts (for more details see for example [7]).

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

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. For example, we can represent the combined relevance of the attributes "name" and "surname" for the category "person".

Preprocessing Before applying the Bayesian Model to our data we performed a two steps of preprocessing. The first step, (*syntactic preprocessing*), involved extracting the terms from the queries. Note that terms could be a single word or a combination of words (compound keywords). Even though we encouraged participants to use quotation marks to point out the compound keywords, the majority of them disregarded this suggestion, improving the need for a subsequent parsing of the queries into terms. Moreover in this phase we 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*). In table 3 we report two examples of the two-step 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	$t_1 \Rightarrow$ event name
	t_2=2008	$t_2 \Rightarrow$ date:year
	t_3=Rome	$t_3 \Rightarrow$ location:city
Q_2= McCain Republican	t_1=McCain	$t_1 \Rightarrow$ surname
	t_2=Republican	$t_2 \Rightarrow$ political party

Table 3. 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 [5]). Over 35% contained only one term and less than 3% of the queries contained five or more terms. Comparing the query length of the refined queries with that of the one-shot queries (i.e., queries without refinement) the average query length increased from 1.78 to 2.56 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 4 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 five attributes with the highest relevance.

Entity Type (e)	Attribute type (a)	$p(e|a)$
Person	surname	0.97
	first name	0.96
	full name	0.96
	affiliation	0.85
	occupation	0.83
Organization	organization name	0.98
	activity	0.85
	organization type	0.85
	part of	0.49
	place:country	0.15
Event	event type	0.97
	event name	0.96
	date:year	0.92
	date:month	0.84
	protagonist:surname	0.79
Artifact	artifact type	0.97
	artifact name	0.94
	brand	0.93
	model	0.91
	features	0.83
Location	location name	0.98
	location type	0.89
	use	0.63
	place:province	0.59
	attraction	0.55

Table 4. Bayesian Relevance: high-level 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 entity target.

However, 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). Note that this table contains two different lists of attributes. The first list corresponds to one-shot queries, the second list is derived from the results of refined queries. We believe that these results could provide valuable insights to study different aspects of query formulation such as refining query strategies.

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 [6]). 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 [17] not uncommon in the context of products [11]). This result reveals another important aspect of the identification process: only a subset of entities are prototypically *namable* entities (e.g. person). Since the ability to use proper names depends critically on our ability to trace individual identity, this evidence suggests interesting research issues. For example, why are people willing to assign proper names only to certain types of entities?

The typology of the entity is another 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.

Events are identified by means of both temporal and spatial attributes whereas locations are preferentially described in terms of spatial attributes. Another common way to specify events is referring to the protagonist of the event (an example is given by the entity type "show"). The identification by author is largely used for some subtypes of artifact such as "artwork" and "book". Qualitative features (e.g. color, gender intended for, material) are extensively used for "clothing". Sometimes location are identified by means of "tourist attractions", such as monuments, famous building or square for city. Other attributes are distinctive of one subtype such as "editor" for book, "type of cuisine" for "restaurant", "number of stars" for "hotel'.

Entity Type (e)	One-shot Queries		Refined Queries	
	Attribute type (a)	$p(e\|a)$	**Attribute type** (a)	$p(e\|a)$
Tourist location	location name	0.66	location name	0.64
	attraction	0.15	location type	0.26
	location type	0.13	attraction	0.22
	tourist agency	0.10	place:island	0.13
			tourist agency	0.10
City	city name	0.67	city name	0.66
	administrative role	0.19	administrative role	0.23
	attraction	0.12	attractions	0.23
	place:country	0.11	famous for	0.11
Shop	shop name	0.53	shop name	0.54
	type of shop	0.21	product type	0.29
	brand	0.18	type of shop	0.22
	product type	0.17	brand	0.14
			ways of purchase	0.10
Hotel	hotel name	0.64	hotel name	0.65
	type of hotel	0.16	place:quarter	0.16
	place:city	0.11	type of hotel	0.13
			number of stars	0.10
Restaurant	restaurant name	0.59	restaurant name	0.57
	type of cuisine	0.43	type of cuisine	0.48
	place:city	0.12	place:address	0.21
	address:street or square	0.11	place:city	0.10
			place:quarter	0.10

Table 5. Bayesian Relevance: Location

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.

3.3 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 2 that shows the probability distribution of the attribute types for the entity type Person. We note, for example, that "first name" is the attribute with the highest probability in first position, whereas "surname" is the preferred attribute in second position.
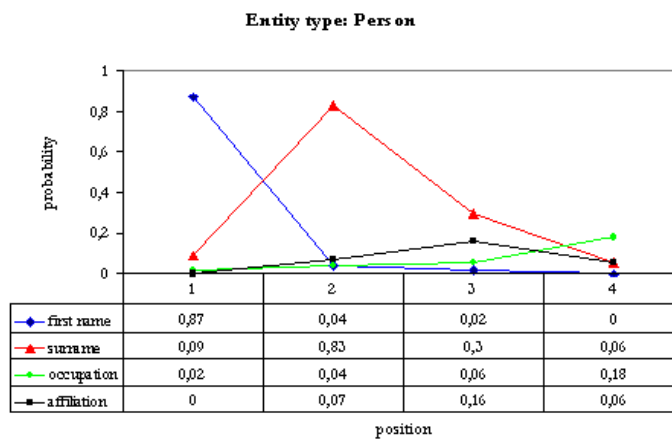
Entity type: Person

	1	2	3	4
first name	0,87	0,04	0,02	0
surname	0,09	0,83	0,3	0,06
occupation	0,02	0,04	0,06	0,18
affiliation	0	0,07	0,16	0,06

Fig. 2. Probability distribution of attribute position

3.4 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 [14] 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.

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

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 can note that the co-occurrence of the "name-surname" pair is very high.

	first name	surname	full name	affiliation	occupation
first name	1	0.84	0	0.06	0.03
surname	0.84	1	0	0.07	0.04
full name	0	0	1	0.08	0.07
affiliation	0.06	0.07	0.08	1	0.05
occupation	0.03	0.04	0.07	0.05	1

Table 6. Co-occurence of attribute types: Person

3.5 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"[5], 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. 2) 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 Q2? Our study can provide some interesting insights. From our data we know that when people search for a city, "city name" is the most relevant attribute (see table 5), as well as the most likely attribute in the first position. From these evidences, we conclude that "Washington" in Q2 is more likely to be city name and finally the entity type of the query is "Location".

---

[5] A fact that is uncomplicated to detect by applying well-known NLP methods such as the use of a thesaurus of person names.

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 of course 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] Barbara Bazzanella, Themis Palpanas, and Heiko 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] Barbara Bazzanella, Heiko Stoermer, and Paolo 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] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio 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] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] B.J. 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.

[6] Saul Kripke. *Naming and Necessity*. Oxford, Basil Blackwell, 1980.

[7] L. Lombardi and G. Sartori. Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, 2007.

[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] G. Sartori, L. Lombardi, and L. Matiuzzi. Semantic relevance best predicts normal and abnormal name retrieval. *Neuropsychologia*, 43:754–770, 2005.

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

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

[13] Heiko Stoermer and Paolo 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.

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

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

[16] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David 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.

[17] Linda Wetzel. Types and tokens. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2008.