FindiLike: Preference Driven Entity Search

Kavita Ganesan
University of Illinois at Urbana Champaign
Urbana, 61801
kganes2@illinois.edu

ChengXiang Zhai
University of Illinois at Urbana Champaign
Urbana, 61801
czhai@cs.uiuc.edu

ABSTRACT

Traditional web search engines enable users to find documents based on topics. However, in finding entities such as restaurants, hotels and products, traditional search engines fail to suffice as users are often interested in finding entities based on structured attributes such as price and brand and unstructured information such as opinions of other web users. In this paper, we showcase a preference driven search system, that enables users to find entities of interest based on a set of structured preferences as well as unstructured opinion preferences. We demonstrate our system in the context of hotel search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Algorithm

1. INTRODUCTION

Web search engines enable users to find all sorts of documents based on a topic of interest. However, with the growth of online content, more and more people are interested in finding entities or objects instead of just documents. This is especially true in decision making scenarios where a user would often like to find entities such as hotels, restaurants and doctors based on their personal requirements. While current search engines like Google are able to recognize certain types of entities (e.g. products and location) these search engines have limited capability in assisting users with decision making. Thus, in a decision making scenario such as choosing a place to eat or a doctor to see, users would turn to sites like Yelp which have better support for decision making where users can select entities of interest based on attributes such as *price*, *location* and *service* and also by reading the unstructured *reviews* of these entities.

Similar to Yelp, vertical search systems like Amazon, Hotels.com, and Bing Shopping facilitate decision making by providing domain specific navigation capabilities in the form of search filters. These filters which are often based on structured information (e.g. *price*, *brand* and *color*), help users to quickly narrow into entities of interest. However, filters based on only structured information, limit the capability for selecting entities based on the unstructured opinions of other users, which is another important factor in decision making. The closest to an 'opinion filter' is the ability to limit entities by the *overall user ratings* which would still force users to read the reviews to ensure that the opinions within these reviews fulfill their requirements. Suppose, a user was looking for a place to eat and wanted a *quiet* restaurant with *good service*. In this case, just limiting the entities by the overall ratings would clearly not be useful. The user would still need to digest the reviews of all restaurants in consideration to find those that satisfy this criteria. Further, not all sites have the 'overall ratings' feature which makes it even harder to leverage existing opinions.

In reality, the opinions of other users is an important influencing factor in our day to day decision making tasks, ranging from which doctor to see to which of the many smart phones to purchase. However, digesting all the available opinions is time consuming and can become confusing over time due to the sheer volume of available opinions. To truly facilitate decision making, opinions should be leveraged in a more efficient manner and should be tightly integrated with the core decision making components of a system.

Existing works [4, 5, 6] have attempted to resolve this problem through summarization of opinions to help users better digest all the opinions. However, when dealing with a large number of entities, even summaries would get confusing as users would still need to keep track of how well each entity fulfills their opinion requirements. Thus, to provide a more direct support for a user's decision making task, we have developed FindiLike, a novel system capable of ranking interesting entities such as hotels based on a set of heterogenous preferences with unstructured *opinion preferences* being a major component of the system. The idea behind FindiLike is to allow users to specify key preferences upfront to the system. These preferences include structured preferences (e.g. *price* and *distance*) as well as unstructured *opinion preferences* that can be specified using descriptive keywords (e.g *clean rooms, cheap, good breakfast* in the context of hotels). With this, the system then scores relevant entities based on how well these entities match the specified preferences. What makes this system unique is that unlike faceted navigation which only filters out 'irrelevant' entities, FindiLike ranks entities by how well *key* preferences are matched, giving users the flexibility in selecting entities based on preference tradeoffs. On top of that, FindiLike allows users to analyze ranked entities using opinion summarization tools which is rarely available with other entity search systems. In the long run, FindiLike aims to evolve into a complete decision making platform for different types of entities. We demonstrate our current system in the context of hotel search. Additional information about this demo can be found at: `http://info.findilike.com`.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

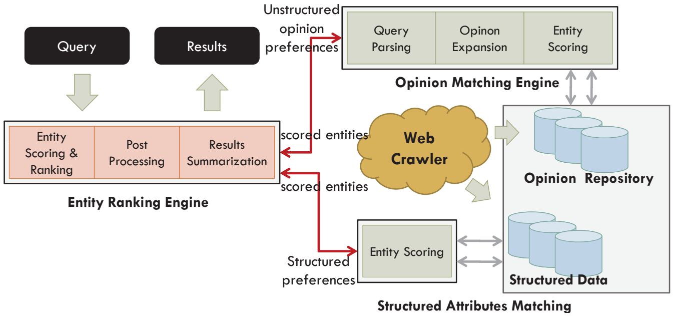


Figure 1: Preference driven entity search architecture. The architecture supports both structured and unstructured preferences.

2. ARCHITECTURE

In this section, we provide a brief overview of the FindiLike system architecture. FindiLike is a web application that enables users to *find* entities based on structured and unstructured set of preferences. Although this ranking task resembles the entity ranking task studied widely by the information retrieval and database communities [7], our task is actually quite different. The goal of entity ranking or entity retrieval is to return relevant entities instead of just documents. The entities are ranked according to how well these entities satisfy a *topic* described in *natural language text*. While the goal of FindiLike is also to retrieve relevant entities, FindiLike ranks entities in the order of likelihood an entity matches a set of *user preferences* rather than just a topic described in plain text. We thus refer to our special set-up as *preference driven entity search*.

Just like any other entity search engines, the FindiLike system consists of several key components ranging from the user query component to a data collection component as shown in Figure 1. In brief, the system takes in user specified preferences and sends these preferences to the relevant scoring engines: opinion preferences to the *opinion matching engine* and structured preferences to *structured attributes matching* module. These scoring engines score a subset of entities (e.g. all hotels in a particular location) based on how well these entities match a given preference. The individual preference scores are then combined and the entities are re-ranked based on these new scores. The summarization module generates a summary of the top N relevant entities which are then displayed to the user. The user then has the option of adding more preferences or has the option of using the analysis tools to further assist them with decision making. In the next few sections, we provide more information about some of the key components of the system.

2.1 User Query

The query to the FindiLike system is a set of preferences. These preferences can be structured by nature such as preference for price, preference for distance and etc. and can also include unstructured preferences for opinions (e.g. desiring a clean hotel when finding hotels at a destination). While opinions can be extracted and used as structured preferences, this information extraction task would be very costly on a large scale and would also force users to express preferences on pre-defined aspects of an entity, which is rather restrictive. With FindiLike, we avoid the need for any information extraction by directly using the review texts of each entity as will be explained in Section 2.2.

The structured preferences provided by the user can vary greatly depending on the application domain. In the case of hotel search, we allow the user to explicitly specify structured preferences on distance from a particular landmark and the desired price range. While various attributes of an entity may be used for preference based ranking, we believe that it is only essential to use the most important attributes for such a ranking feature leaving the rest of the attributes as filters just as in faceted navigation.

As for the unstructured opinion preferences, we ask users to state their opinion preferences using a set of descriptive keywords. These keywords would indicate what the user desires in the different aspects of an entity. For example, to show desire for clean hotels with friendly staff, the user may specify a query such as *clean rooms, friendly staff* or *clean place, friendly service*. The ability to specify preferences using free-form text enables users to express preferences on any arbitrary aspect and for any type of desired opinions. In accepting a user's unstructured preferences, different types of user interfaces may be used. The most general interface would be a single text field that would allow users to express preferences using natural keywords. Aspects in the query can then be obtained using various query segmentation techniques. To make this more practical, in our system, users can specify all their preferences in a single query box using a special delimiter such as 'and' or comma to separate each preference. We also allow users to incrementally add preferences as needed instead of re-entering the entire query.

This type of unstructured expression for opinion related preferences, brings about a new type of query understanding problem. The opinion preferences expressed by users can often be ambiguous, and there can be multiple ways to express similar preferences. For example, the expression "good breakfast" is similar in meaning to "great breakfast". To help with the matching of opinions, it would thus be beneficial to expand such a query by adding synonyms of the sentiment word. To this end, we have implemented ideas from [1] in dealing with some of the query understanding problems namely for the task of opinion expansion.

2.2 Entity Ranking Engine

For a given class of entities (e.g. hotels in Chicago), the entity ranking engine takes in a set of preferences and attempts to find entities that match all of these preferences. Each entity in the given class is scored based on how well it matches each preference and then the scores are combined. The top N scoring entities are returned as relevant results. By default, these entities are ranked in the order of likelihood an entity matches a user's preferences. More formally, given a set of preferences, $P = \{p_1, ...p_n\}$, the score of an entity E from class i is computed as follows:

$$S(P, E_i) = \frac{1}{n}[S_{structured}(P, E_i) + S_{opinions}(P, E_i)]$$

where, $S_{structured}(P, E_i)$ is scoring based on structured preferences such as price and distance and $S_{opinions}(P, E_i)$ is scoring based unstructured preferences which in this case is opinions. All preference scorings are on a scale of 5.

2.2.1 Scoring of Structured Preferences

The scoring of structured preferences is based on how well an entity fulfills the given criteria. Suppose, the preference for the *maximum price* aspect is set at $60. If an entity's price is $70, this entity does not quite fulfill the given criteria. In such a case, instead of completely penalizing this entity, the entity is given a score lower than the maximum possible score depending on how much it violates the criteria. In this example, the entity's price exceeds the maximum price by $10, so this entity may be assigned a score of 4/5 instead of the maximum score of 5/5 for this specific aspect. On the other hand, if the entity's price falls within the maximum price requirement, then the entity immediately receives a full score on this aspect.There are several advantages to scoring entities in this way as opposed to completely eliminating entities. First, a user's requirement can sometimes be unrealistic and complete elimination could yield in no results being returned, which is not good from the perspective of user experience. In contrast, scoring entities with respect to the level of violation would yield in results that most closely match the specified criteria only that the scores could be much lower, which would then encourage users to change their expectations. Users may also be willing to loosen their requirements on some aspects if other aspects of an entity is in their favor. Using the previous example, while the entity's price exceeds the maximum desired by $10, this entity may have matched other preferences extremely well. In this case, the user may decide to give in to the higher price as all other aspects of the entity are appealing to the user.

2.2.2 *Scoring of Opinion Preferences*

As we avoid the need for costly opinion mining and information extraction, the scoring of opinion preferences differs from how structured preferences are scored. In FindiLike, opinion preferences expressed using descriptive keywords are scored against the review texts written by both experts and average users. This matching task is quite different from keyword search in databases [8] where the goal is to find objects where any of its fields match the given keywords. Our idea is to represent each entity with the unstructured text of all the reviews of that particular entity, often available from various websites. Given a user's keyword preferences that expresses the desired features of an entity (e.g *clean and safe* for a hotel), we then score the relevant entities based on how well its reviews match the user's preferences using a retrieval model as described in Ganesan & Zhai [1]. The more relevant mentions there are in the reviews of an entity, the higher the score an entity receives. Since opinions are highly subjective, it is often difficult to clearly determine if the opinions accurately describe an entity. However, if different users express similar dissatisfaction and appreciation about an entity, then its more likely that these opinions reflect accurate descriptions of the entity which is the idea behind our scoring mechanism.

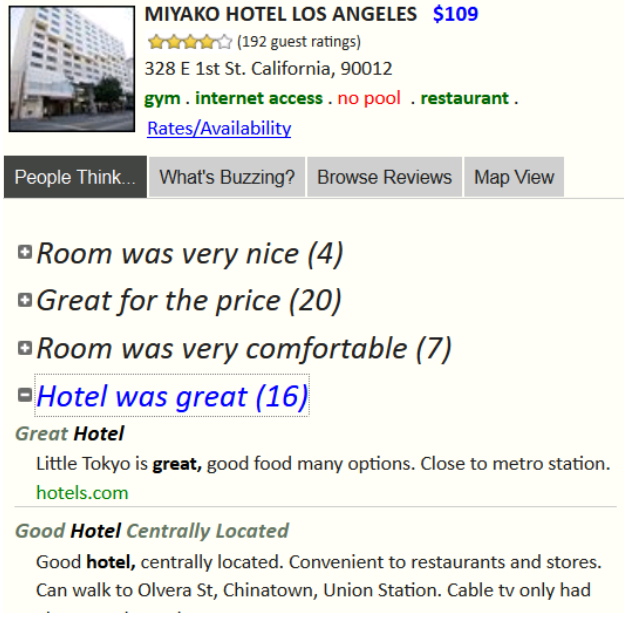
3. DEMO

We will demonstrate our system in the context of hotel search which is accessible at `http://www.findilike.com`. We will demonstrate the following features of FindiLike:

Ranking hotels by preferences. FindiLike enables users to find hotels using opinion driven preferences and other preferences such as distance and price. Suppose the user needs to find hotels close to the Los Angeles Convention Center and wants a hotel in a safe location. Using the FindiLike system, the user can find relevant hotels based on all these requirements. The preference for proximity to the Convention Center can be specified under the 'distance' tab and the preference for hotels that are said to be safe can be specified using the main search box using natural keywords such as *safe neighborhood.* Once all the requirements have been specified, all hotels in the Los Angeles area are then scored based on how well each hotel matches the specified preferences. The results are then ranked in the descending order of the scores as shown in Figure 2 (under 'YourMatch'). The individual preference scores are displayed using 'green stars'.

Summarizing ranked hotels. In traditional web search, a user often navigates into search results based on the relevance of the summary snippets to the query. Since, in our case, the user is looking for a hotel based on a set of preferences, a summary of the selected preferences is displayed to the user (as shown in Figure 2). For opinion preferences, snippets from the user reviews are displayed. For the distance preference, distance of the hotel from the selected landmark along with the total driving time are displayed.

Browsing opinions via summaries. To further assist users in their decision making process, we help users navigate the opinion space using automatically generated summaries. In most systems, the closest to an opinion summary is the averaged overall ratings provided by different users. Unstructured summaries which can often be more informative [2, 3], is almost never available in existing systems. FindiLike is capable of generating unstructured summaries of opinions, so as to help users digest the most common praises or complaints within the reviews. A snapshot of summaries generated for a hotel in Los Angeles is shown in Figure 3. In addition to summaries, to help users visualize mentions within opinions, we also provide a tag cloud representation of the common mentions under the 'What's Buzzing' tab as shown in Figure 4.



MIYAKO HOTEL LOS ANGELES $109

★★★★☆ (192 guest ratings)

328 E 1st St. California, 90012

gym . internet access . no pool . restaurant .

⊞ *Room was very nice (4)*

⊞ *Great for the price (20)*

⊞ *Room was very comfortable (7)*

⊟ *Hotel was great (16)*

*Great **Hotel***

Little Tokyo is **great,** good food many options. Close to metro station.

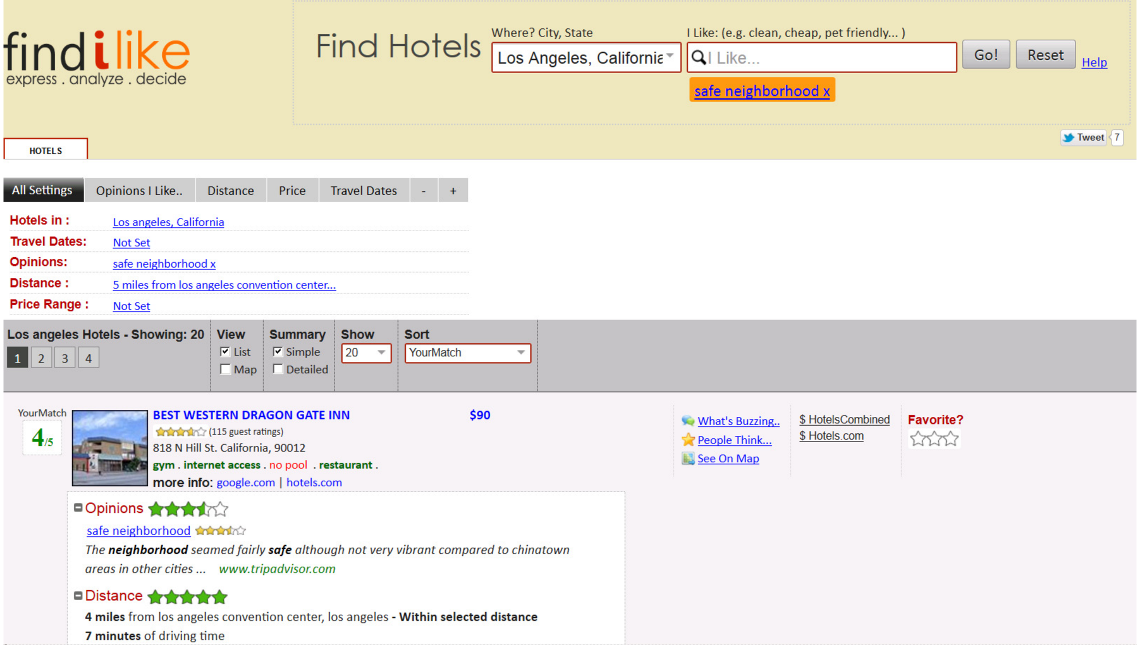
hotels.com

*Good **Hotel** Centrally Located*

Good **hotel,** centrally located. Convenient to restaurants and stores. Can walk to Olvera St, Chinatown, Union Station. Cable tv only had

Figure 3: Opinion summaries generated by FindiLike for a hotel in Los Angeles. Numbers within parentheses indicate the number of supporting mentions for that particular summary. A click on each summary will display all the supporting reviews.

List view vs. map view. Two kinds of presentational views are supported by the system. The list view as shown in Figure 2 and a map view as shown in Figure 5. In list view, the results are organized in a flat list ranked by the overall preference score. This view provides a detail summary of the results and enables users to easily select links and navigate into other components of the system. In map view, the results are displayed on a Google Map, with a small list type summary of the search results on the left. Individual markers need to be selected to see detailed summaries and to navigate into other components of the system.



Where? City, State: Los Angeles, California

I Like: (e.g. clean, cheap, pet friendly...): I Like...

safe neighborhood x

Hotels in :	Los angeles, California
Travel Dates:	Not Set
Opinions:	safe neighborhood x
Distance :	5 miles from los angeles convention center...
Price Range :	Not Set

Los angeles Hotels - Showing: 20 · 1 2 3 4

View: ☑ List ☐ Map · Summary: ☑ Simple ☐ Detailed · Show: 20 · Sort: YourMatch

YourMatch 4/5

BEST WESTERN DRAGON GATE INN $90
★★★★☆ (115 guest ratings)
818 N Hill St. California, 90012
gym . internet access . no pool . restaurant .
more info: google.com | hotels.com

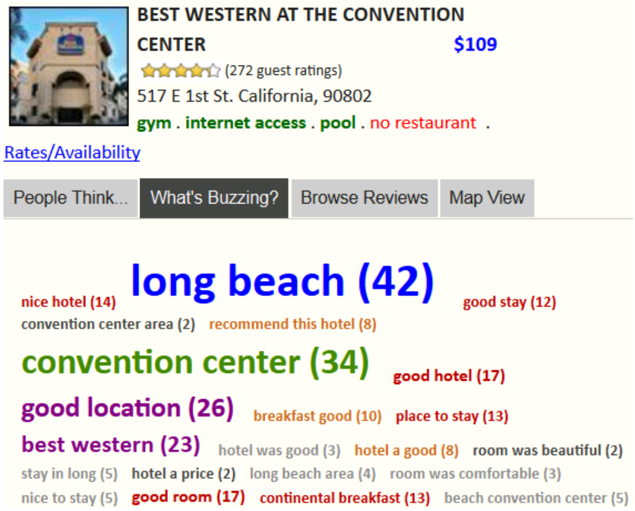
$ HotelsCombined
$ Hotels.com

Favorite? ☆☆☆

⊟ Opinions ★★★½☆
safe neighborhood ★★★½☆
*The **neighborhood** seamed fairly **safe** although not very vibrant compared to chinatown areas in other cities ...* www.tripadvisor.com

⊟ Distance ★★★★★
4 miles from los angeles convention center, los angeles - **Within selected distance**
7 minutes of driving time

Figure 2: FindiLike preference based search. Results shown for the query: "Hotels within *5 miles of the Los Angeles Convention Center* and located in a *safe neighborhood*". *YourMatch* score on far left shows how well all preferences are matched on a scale of 5. Green stars indicate individual preference match levels.



BEST WESTERN AT THE CONVENTION CENTER $109
★★★★☆ (272 guest ratings)
517 E 1st St. California, 90802
gym . internet access . pool . no restaurant .

Rates/Availability

long beach (42) · nice hotel (14) · good stay (12) · convention center area (2) · recommend this hotel (8) · convention center (34) · good hotel (17) · good location (26) · breakfast good (10) · place to stay (13) · best western (23) · hotel was good (3) · hotel a good (8) · room was beautiful (2) · stay in long (5) · hotel a price (2) · long beach area (4) · room was comfortable (3) · nice to stay (5) · good room (17) · continental breakfast (13) · beach convention center (5)

Figure 4: 'Clickable' tag cloud visualization of common mentions within reviews.

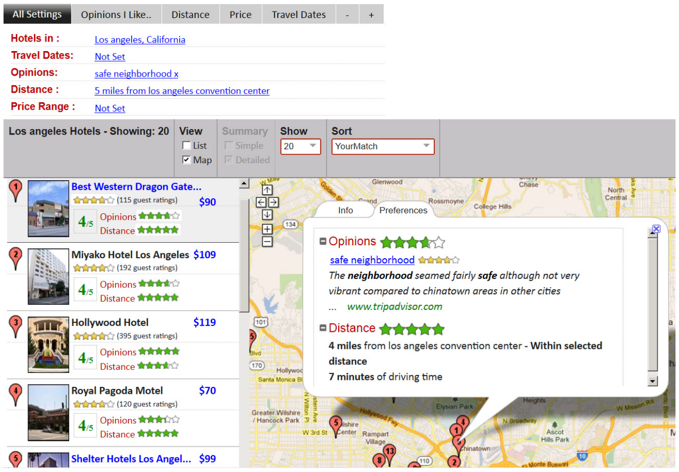
4. REFERENCES

[1] K. Ganesan and C. Zhai. Opinion-based entity ranking. *Information Retrieval*, 2011.

[2] K. Ganesan, C. Zhai, and J. Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, 2010.

[3] K. Ganesan, C. Zhai, and E. Viegas. Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*, 2012.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.



Hotels in :	Los angeles, California
Travel Dates:	Not Set
Opinions:	safe neighborhood x
Distance :	5 miles from los angeles convention center
Price Range :	Not Set

Los angeles Hotels - Showing: 20 · View: ☐ List ☑ Map · Summary: ☐ Simple ☑ Detailed · Show: 20 · Sort: YourMatch

1. Best Western Dragon Gate... ★★★★☆ (115 guest ratings) $90 · 4/5 · Opinions ★★★★☆ · Distance ★★★★★
2. Miyako Hotel Los Angeles $109 ★★★★☆ (192 guest ratings) · 4/5 · Opinions ★★★½☆ · Distance ★★★★★
3. Hollywood Hotel $119 ★★★★☆ (395 guest ratings) · 4/5 · Opinions ★★★★★ · Distance ★★★½☆
4. Royal Pagoda Motel $70 ★★★★☆ (120 guest ratings) · 4/5 · Opinions ★★★½☆ · Distance ★★★★★
5. Shelter Hotels Los Angel... $99

⊟ Opinions ★★★½☆
safe neighborhood ★★★½☆
*The **neighborhood** seamed fairly **safe** although not very vibrant compared to chinatown areas in other cities ...* www.tripadvisor.com

⊟ Distance ★★★★★
4 miles from los angeles convention center - **Within selected distance**
7 minutes of driving time

Figure 5: Results in map view. Red markers represent relevant results. Yellow marker represents the selected landmark.

[5] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140, Madrid, Spain, 2009. ACM.

[6] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. *In Proceeddings of HLT-NAACL '07*, pages 300–307, 2007.

[7] A. P. Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Focused access to xml documents. chapter Overview of the INEX 2007 Entity Ranking Track, pages 245–251. Springer-Verlag, Berlin, Heidelberg, 2008.

[8] J. X. Yu, L. Qin, and L. Chang. *Keyword Search in Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2010.