MASFA: Mass-collaborative Faceted Search for Online Communities

Seth B. Cleveland
Texas State University
San Marcos, TX, 78666, USA
sc1439@txstate.edu

Byron J. Gao
Texas State University
San Marcos, TX, 78666, USA
bgao@txstate.edu

ABSTRACT

Faceted search combines faceted navigation with direct keyword search, providing exploratory search capacities allowing progressive query refinement. It has become the de facto standard for e-commerce and product-related websites such as amazon.com and ebay.com. However, faceted search has not been effectively incorporated into non-commercial online community portals such as craigslist.org. This is mainly because unlike keyword search, faceted search systems require metadata that constantly evolve, making them very costly to build and maintain. In this paper, we propose a framework MASFA that utilizes a set of non-domain-specific techniques to build and maintain effective, portable, and cost-free faceted search systems in a mass-collaborative manner. We have implemented and deployed the framework on selected categories of Craigslist to demonstrate its utility.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Search process*

Keywords: faceted search; faceted navigation; taxonomy; interactive information retrieval; human computation; mass-collaboration; crowdsoursing; named entity recognition

1. INTRODUCTION

Faceted search adds structured browsing, or faceted navigation, to direct keyword search, supporting interactive and progressive query refinement [6]. It well addresses weaknesses of conventional discovery-oriented search paradigms and has emerged as a foundation for interactive information retrieval. User studies demonstrate that faceted search interfaces are intuitive and easy to use, providing more effective information-seeking support than conventional search paradigms [2, 4, 6]. Faceted search has become increasingly prevalent in online information access systems and is currently the de facto standard for e-commerce and product-related websites, such as amazon.com, ebay.com, walmart.com, bestbuy.com, homedepot.com, and carmax.com.

Challenges for non-commercial sectors. Despite the prominent success in e-commerce, faceted search has not been effectively incorporated into non-commercial online community portals such as craigslist.org and medhelp.org. This is mainly because comparing to keyword search, faceted search systems require structured metadata while community data are mainly free texts and unstructured. In addition, such metadata constantly evolve following the life cycles of products or topics. Thus faceted search systems are costly to build and maintain. Community portals are usually not-for-profit and cannot afford to hire and train employees as e-commerce businesses. Existing automatic faceted search techniques do not port well to new domains. Facets are domain-specific. The facet structures organizing cars are different from the ones organizing clothes. Existing techniques generally assume domain knowledge of facet structures and make use of domain-specific, hand-crafted rules or machine learning models, which are costly to generate and update, and not portable across domains [3].

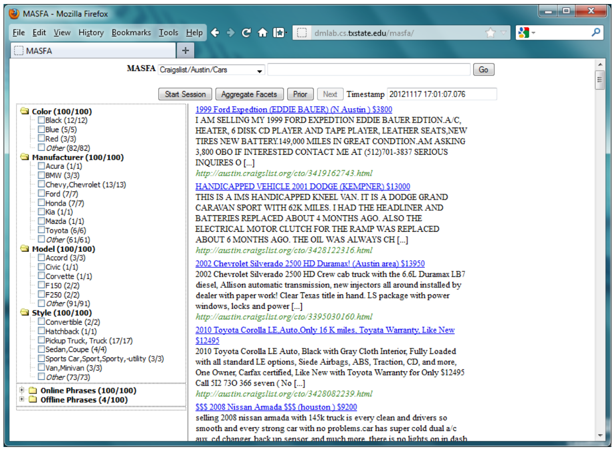
Today, thriving online communities have re-defined modern society and transformed the way day-to-day activities are conducted. People spend more and more time participating in various online communities on a daily basis. For example, craigslist.org accounts for nearly 2% of global internet traffic. Despite the economic and technical challenges, there is an increasing need to facilitate faceted search for online communities, enabling more effective use and management of community data.

Our approach. In this paper, we explore a novel direction in enabling faceted search for online communities, utilizing the power of mass-collaboration or crowdsourcing [1]. In particular, we introduce MASFA, the first framework for mass-collaborative faceted search that can be deployed and operated free of cost. MASFA takes a human-machine partnership approach, where humans, i.e., community members, contribute to the faceted search system while using it, and machines assist humans in this process based on a set of non-domain-specific techniques. The MASFA approach is completely portable and can be deployed and maintained in any application domain in a cost-free manner. Yet it can be highly effective, significantly reducing user search time.

A demonstrating prototype (dmlab.cs.txstate.edu/masfa/) of MASFA has been implemented and deployed for selected categories of Craigslist based on Craigslist RSS feed. The prototype is open to public access and Figure 1 shows a screenshot of it. The left-hand panel presents a set of facets, i.e., taxonomies. The right-hand panel presents refined search results (Craigslist posts) for a given query that satisfy the condition specified by the selected facet values. A community member can edit the facets by adding, deleting, and modifying facet names and values.

In MASFA, community members can arbitrarily edit the facets. Such edits are recorded in a temporal database [5], so that the facets can be brought back to any previous ver-

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.


MASFA - Mozilla Firefox
File Edit View History Bookmarks Tools Help
dmlab.cs.txstate.edu/masfa/
MASFA
+
Go
Start Session
Aggregate Facets
Prior
Next

MASFA Craigslist/Austin/Cars

Timestamp 20121117 17:01:07.076

Color (100/100)
- □ Black (12/12)
- □ Blue (5/5)
- □ Red (3/3)
- □ *Other* (82/82)

Manufacturer (100/100)
- □ Acura (1/1)
- □ BMW (3/3)
- □ Chevy,Chevrolet (13/13)
- □ Ford (7/7)
- □ Honda (7/7)
- □ Kia (1/1)
- □ Mazda (1/1)
- □ Toyota (6/6)
- □ *Other* (61/61)

Model (100/100)
- □ Accord (3/3)
- □ Civic (1/1)
- □ Corvette (1/1)
- □ F150 (2/2)
- □ F250 (2/2)
- □ *Other* (91/91)

Style (100/100)
- □ Convertible (2/2)
- □ Hatchback (1/1)
- □ Pickup Truck, Truck (17/17)
- □ Sedan,Coupe (4/4)
- □ Sports Car,Sport,Sporty,-utility (3/3)
- □ Van,Minivan (3/3)
- □ *Other* (73/73)

Online Phrases (100/100)
Offline Phrases (4/100)

1999 Ford Expedtion (EDDIE BAUER) (N Austin) $3800
I AM SELLING MY 1999 FORD EXPEDTION EDDIE BAUER EDTION.A/C, HEATER, 6 DISK CD PLAYER AND TAPE PLAYER, LEATHER SEATS,NEW TIRES NEW BATTERY.149,000 MILES IN GREAT CONDTION.AM ASKING 3,800 OBO IF INTERESTED CONTACT ME AT (512)701-3837 SERIOUS INQUIRES O [...]
http://austin.craigslist.org/cto/3419162743.html

HANDICAPPED VEHICLE 2001 DODGE (KEMPNER) $13000
THIS IS A IMS HANDICAPPED KNEEL VAN. IT IS A DODGE GRAND CARAVAN SPORT WITH 62K MILES. I HAD THE HEADLINER AND BATTERIES REPLACED ABOUT 4 MONTHS AGO. ALSO THE ELECTRICAL MOTOR CLUTCH FOR THE RAMP WAS REPLACED ABOUT 6 MONTHS AGO. THE OIL WAS ALWAYS CH [...]
http://austin.craigslist.org/cto/3428122316.html

2002 Chevrolet Silverado 2500 HD Duramax! (Austin area) $13950
2002 Chevrolet Silverado 2500 HD Crew cab truck with the 6.6L Duramax LB7 diesel, Allison automatic transmission, new injectors all around installed by dealer with paper work! Clear Texas title in hand. LS package with power windows, locks and power [...]
http://austin.craigslist.org/cto/3395030160.html

2010 Toyota Corolla LE.Auto.Only 16 K miles. Toyata Warranty. Like New $12495
2010 Toyota Corolla LE Auto, Black with Gray Cloth Interior, Fully Loaded with all standard LE options, Siede Airbags, ABS, Traction, CD, and more, One Owner, Carfax certified, Like New with Toyota Warranty for Only $12495 Call 5I2 73O 366 seven (No [...]
http://austin.craigslist.org/cto/3428082239.html

$$$ 2008 Nissan Armada $$$ (houston) $9200
selling 2008 nissan armada with 145k truck is every clean and drivers so smooth and every strong car with no problems.car has super cold dual a/c aux, cd changer, back up sensor, and much more. there is no lights on in dash

Figure 1: Screenshot of MASFA.

sion for a given timestamp. On the other hand, machines collect historical data and generate frequent phrases, which can be used to suggest addition or removal of facet values. Machines also contribute to the formation of a robust, aggregated version of facets from the numerous human-edited versions based on their life span and usage statistics. The aggregation incorporates clustering techniques and is expected to smooth out noise and turbulence that are common in crowdsourcing tasks.

2. THE MASFA FRAMEWORK

2.1 Overview

Figure 2 shows the main architecture of MASFA. For each category of data source, e.g., Cars, MASFA maintains a set of facets that evolve over time. For a given keyword query q within a selected category, the Query Processing module produces a set R of search results. Throughout the paper, search results are often referred to as *items* that can be product descriptions or Craigslist posts. Then, based on a chosen version of facets F, the Faceted Navigation module allows the user to interactively and progressively refine the search results and produce R', a refined set of items.

The Facet Editing and Management module takes human and machine efforts to build and maintain facets. A community member (user) can start an editing session by clicking the *Start Session* button. Then s/he can edit the facets by adding, deleting, and modifying facet names and values. A successful editing session will result in a new version to be created with an assigned timestamp. The module has a temporal database back end that records the entire evolving history of facets. A temporal database [5] is a database with built-in time aspects that is able to store different database states. In MASFA, every version can be brought back by specifying its associated timestamp in the *Timestamp* editbox. There are also *Prior* and *Next* buttons that can be used to navigate through all the recorded versions. By clicking the *Aggregated Facets* button, users can obtain a synthetic set of facets that are robust against anomilies.

2.2 Interface and Semantics

In a typical faceted search interface, there are a set of facets or taxonomies. Each facet has a *name*, and is associated with *facet values* that are *exhaustive* (collectively covering all the items) and *mutually exclusive* (not covering any item in common). For example, facet *Make* would have *Toyota*, *Chevrolet* ... as its facet values. The values within a facet can be a flat or hierarchical list.

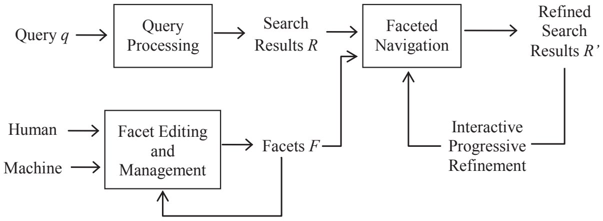

Query q
Query Processing
Search Results R
Faceted Navigation
Refined Search Results R'
Human
Machine
Facet Editing and Management
Facets F
Interactive Progressive Refinement

Figure 2: Architecture of MASFA.

Flat list structure. In MASFA, facet values appear as a flat list. This is a design, not a technical, option. In general, faceted search works better with a broad taxonomy that is relatively shallow, as this lets users combine more perspectives rather than get stuck in an eternal drill down, which causes fatigue (www.uie.com/articles/faceted_search/). Many commercial sites such as Linkedin people search and the Costco wireless phone shopping site (membershipwireless.com /index.cfm) use flat lists for clarity of interface and logic. MASFA implements a mass-collaborative framework, where it is particularly important to avoid unnecessary confusions and complications, making sure that contributors share the same or similar understanding about the system and have a common ground to work on collaboratively. Flat lists are much easier to visualize, comprehend, and edit.

Relaxation of exhaustiveness and mutual exclusiveness. While the conventional exhaustiveness and mutual exclusiveness constrains provide clear classification of items, their enforcement would incur significant difficulty for community members to construct facets. MASFA relaxes these constraints by allowing incomplete and overlapping coverage of items. A special value *Other* is added whenever necessary to collect the items not covered by the sibling values within a facet. Reasonable overlapping of items will not be purposely ruled out. For example, if a car has both black and blue colors, then the corresponding car post would be covered by both *Black* and *Blue* values under the facet *Color*. In practice, a well-behaved MASFA faceted search system would be nearly exhaustive and nearly mutually exclusive. This relaxation does not compromise the utility of the system much, yet successfully avoiding significant building and maintenance costs.

Implicit metadata generation. Unlike conventional faceted search, a facet value in MASFA is not a single value, but a set of positive and negative phrases separated by commas. It represents a Boolean formula and covers the items that satisfy the formula. Let V be a facet value, where $P_1, P_2, \cdots, P_m$ constitute the set of positive phrases and $-N_1, -N_2, \cdots, -N_n$ constitute the set of negative phrases. Then V corresponds to a Boolean formula of $(P_1 \vee P_2 \vee \cdots \vee P_m) \wedge \neg(N_1 \vee N_2 \vee \cdots \vee N_n)$. The items satisfy the formula are the ones that contain any of the positive phrases and do not contain any of the negative phrases. This interpretation corresponds to an implicit way of generating metadata, where the satisfying items are "labeled" by a set of positive phrases $P_1, P_2, \cdots, P_m$ and a set of negative phrases

$-N_1, -N_2, \cdots, -N_n$. Note that this implicit named entity recognition and classification mechanism is not domain-specific. It can be utilized in any application domain and does not incur maintenance or update cost.

In practice, the positive phrases are usually different mentions of the same (or similar) target feature, for example, *Chevrolet* and *Chevy*. The negative phrases are used to weed out different features that happen to have the same or similar mentions to those of the target feature. For example, if the target feature is sport cars (small cars designed for performance), then we may want to weed out sport utility vehicles (special purpose vehicles for towing with on and off road capabilities) by using a negative phrase *-sport utility*.

Selected facet values. During faceted navigation, multiple facet values maybe selected from multiple facets. MASFA implements the CNF semantics for the selected facets, where they form a conjunction of disjunctions. For example, if V_1 (e.g., *Ford*) and V_2 (e.g., *Honda*) are selected from facet F_1 (e.g., *Manufacturer*) and U_1 (e.g., *Black*) and U_2 (e.g., *Blue*) are selected from facet F_2 (e.g., *Color*), then the compound Boolean formula will be $(V_1 \vee V_2) \wedge (U_1 \vee U_2)$ (e.g., cars made either by Ford or Honda that are either black or blue in color). The refined search results R' will contain all the items from R (the original search results for query q) that satisfy the compound formula.

Item counts. Item count numbers contain important information for progressive query refinement, providing a preview of the refined search results before a facet value is actually selected. In MASFA, each facet value V is associated with an item count in the form of x/y, where y is the total number of original results for query q that are covered by V. The y number is a function of q and V and will not change throughout the progressive refinement process for query q.

If a sibling value V' within the same facet has been selected, then x indicates the maximum (not exact, because MASFA allows overlapping coverage among sibling facet values) number of the items that can possibly be added (removed) to the refined results if V is selected (de-selected). This is because MASFA implements CNF semantics for selected facet values and selected sibling facet values are OR-connected. Suppose under the facet *Color*, *Black* (9/9) has been selected and there are 9 items in the set of refined results. The subsequent selection of *Blue* (5/5) would add at most 5 items to the refined results if there is no overlapping between *Black* and *Blue*. If one car is black and blue (containing both words in the post), then only 4 items will be added to the refined results.

If none of the sibling values of V has been selected, then x indicates exactly the number of items that will appear in the refined search results if V is selected. This is the case even when some facet values from other facets have been selected because MASFA implements CNF semantics for selected facet values and selected values from different facets are AND-connected. If none of the sibling values but V has been selected, then x only indicates the current number of refined results covered by V (which is also the total number of refined results since V is the only value selected within the facet) and cannot be used to predict the change of number of refined results once V is de-selected. This is due to a convenient but incorrect convention in faceted search: it is considered *all* values within a facet are selected if *none* of them is selected.

Each facet name in MASFA is also associated with an item count in the form of x'/y', where y' indicates the total number of original search results for the initial keyword query q, and x' indicates the total number of refined results for the selected facet values. Obviously, all the facets will share the same x'/y' at all times. Such numbers provide summative information for the progressively refined search results.

2.3 Facet Editing and Management

Facet editing and versioning. A community member can start an editing session and edit the facets by adding, deleting or modifying facet names and facet values. The refined results as well as item counts will be updated immediately and automatically after each edit. No login is required. All edits in MASFA are available through context menus. The editing session will expire in certain period of time (10 minutes) unless renewed. A successful editing session with valid edits will result in a new *version* of facets to be created.

Machine-extracted phrases (Section 2.4) can assist with human editing. For example, phrases *Toyota*, *Honda*, *Audi*, and *BMW* can be moved into one facet labeled *Make*. Phrases *Wagons*, *Convertibles*, and *Pickup Trucks* can be moved into one facet labeled *Vehicle Type*. The machine can also signal out-of-dated facet values suggesting for removal.

User edits are valuable contributions. It is important to keep the historical edits, instead of only the current version of facets, for multiple beneficial purposes such as aggregating user contributions and personalizing user preferences. In MASFA, a temporal database [5] is used to store all the user edits, where addition and deletion (a modification is equivalent to a deletion plus an addition) timestamps are recorded. Specifically, the facet trees are decomposed into pairs of labels that correspond to edges, and the pairs are the actual stored database objects. Pairs, Addition timestamps and Deletion timestamps together form a composite key in a relational table.

Our current prototype does not stress concurrency control. It implements a simple policy that only one user can edit the facets at a session. Community members can edit any version of facets, not necessarily the current one.

Facet aggregation. Given the open nature of crowdsourcing systems, noise and turbulence are common due to differences in preferences and understanding, unintentional execution errors, or malicious spamming. Statistics-based aggregation can effectively smooth out such noise and turbulence. In MASFA, by clicking the *Aggregated Facets* button, users can obtain a synthetic set of facets. The aggregation utilizes clustering techniques on all facets of various human-edited versions. Each cluster represents an equivalence class consisting of different versions of the same facet. For each cluster, the best member is chosen.

The clustering algorithm first decomposes the facets of all versions and ranks them. In general, facets with longer life spans and more usage tend to be more robust, enduring, and popular, and they are ranked high. Then, we process the ranked facets one at a time. If its similarity with an existing cluster (w.r.t. the closest member in the cluster) is big enough, it will be inserted into that cluster. Otherwise, a new cluster will be created.

Pairwise similarity of facets is computed using modified Jaccard coefficient. Let F_1 and F_2 denote two facets, each consisting of a set of facet values, then $similarity(F_1, F_2) = \frac{|F_1 \cap F_2|}{min(|F_1|,|F_2|)}$. We use $min(|F_1|, |F_2|)$ to replace $|F_1 \cup F_2|$ as in the standard Jaccard coefficient to boost the similarity measure. This is because in practice, different facets (e.g., *Color* and *Make*) rarely share some facet values in common. In case two facets do share a few in common, it would be a strong indication that they are actually different versions of the same facet and should be clustered together. The cut-off threshold is set to 10% in MASFA but it is tunable.

2.4 Frequent Phrases

Frequent phrases can be used to suggest addition or removal of facet values, serving as building blocks in facet construction and organization and reducing editing workload of community members. It can also be used to add a layer of machine supervision to reduce turbulence. The extracted phrases can be considered as a superset of the common facet values. They are mixed, not organized into facets, but ranked according to frequency and made clickable, which makes them useful even in query refinement.

In a centralized category of documents (Car in Craigslist), facet values (*Toyota*, *Chevy*, *Blue*, and *Power Window*) are frequently used as feature descriptors. Thus, potentially a frequency-based, non-domain-specific approach can be used to extract such facet values. In MASFA, a category of documents are collected and pre-processed. Then, a suffix-tree-based algorithm is used to extract frequent syntactic phrases. Then, simple cleansing heuristics are applied to remove noisy phrases. The phrase extraction techniques in MASFA are completely non-domain-specific and portable. MASFA performs phrase extraction periodically to identify emerging frequent phrases. We omit the details here.

3. IMPLEMENTATION

Back end. The MASFA back end provides a Craigslist client, facet data and the business logic to manage data. The back end is implemented using the Spring Model View Controller package (MVC). MVC provides custom presentations, JSON serialization, database access, and custom business logic. The back end serves all facets and processed data to the front end as JSON using AJAX (www.json.org).

Craigslist provides data as Really Simple Syndication feeds (RSS) (www.craigslist.org/about/rss). MASFA implements a client that requests and parses RSS feeds using the open source library Rome (rometools.org). From each craigslist result the text fields are scanned into tokens using Jflex version 1.4.3 (jflex.de). Then, each token is normalized by applying a case filter, a stemming filter (snowball.tartarus.org), and a stop word filter to extract terms. Then, an inverted index is created with the extracted terms. The MASFA back end has a Java H2 Database engine (www.h2database.com/html/main.html) that implements a temporal database.

Front end. The MASFA front end provides dynamic user interface for editing facet trees using Javascript and HTML. It allows the user to retrieve facets from the back end for a given timestamp, as well as processed Craigslist data for a given query. The processed Craigslist data is cached to facilitate dynamic updates to facet names, values and item counts. Each user edit is sent to the back end for storage.

jQuery (jquery.com) provides the core functionality for dynamic user interface based on plugins and AJAX queries. The tree is presented using the Dynatree jQuery plugin (code.google.com/p/dynatree). Dynatree provides an interface to dynamically build and manipulate trees in a Web browser. We used the hierarchical selection and checkbox feature of the plugin. A context menu jQuery plugin (abeautifulsite.net/2008/09/jquery-context-menu-plugin) provides a user interface to support creation, update, and deletion of facet names and values as tree nodes.

4. DEMONSTRATION

The MASFA prototype (dmlab.cs.txstate.edu/masfa/) is maintained on a server with two Intel Xeon X5675 processors each having 6 cores @ 3.07GHz, 24GB memory, and 1.3TB disk storage, running Apache Tomcat 6.0.26.

Basic functionalities. The default version of facets is the *current* one. You can issue keyword queries for a chosen category, e.g., Cars. The search results will be organized by the current facets. You may progressively refine your query by selecting labels from multiple facets. Observe that how item counts and refined results are updated when a value is selected. For example, when you select *Toyota*, *Chevy* and *Blue*, the refined results will contain all items describing blue Toyota or blue Chevy cars.

Temporal queries. You may specify a preferred timestamp, and the system will bring back that version of facets to re-organize your search results. You may also click the *Prior* or *Next* buttons to navigate through the stable versions in history, or click the *Aggregated Facets* button to use the aggregated facets.

Editing facets. You may start editing and contributing by clicking the *Start Session* button. You may choose any version, usually the current one, to begin with. You may add, delete and modify any facet names or facet values. All edits are available through context menus. Type Enter to finish an edit. The refined results and item counts will be updated immediately. You may choose a new category for which the facets are not constructed yet. Use the machine-extracted phrases to help you with the construction of facets.

Negative phrases. Negative phrases are a novel feature of MASFA that increases the expressiveness and flexibility of the system. For example, you may create a value -*White* so as to find cars that are not white. You may also modify a value *Blue* to *Blue, -Dark Blue* to exclude dark blue cars from all the blue ones.

5. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under grant No. CNS-1058724 and OCI-1062439.

6. REFERENCES

[1] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, 2011.

[2] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Commun. ACM*, 45:42–49, 2002.

[3] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[4] G. M. Sacco and Y. Tzitzikas. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience.* Springer Publishing Company, Incorporated, 1st edition, 2009.

[5] R. T. Snodgrass. *Developing Time-Oriented Database Applications in SQL.* Morgan Kaufmann Publishers, 1999.

[6] D. Tunkelang. *Faceted Search.* Morgan & Claypool Publishers, 2009.