International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 15th Jan 2019. Link :

http://www.ijiemr.org/main/index.php?vol=Volume-08&issue=ISSUE-01

Title: **INFORMATION RETRIEVAL FOR MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS**

Volume 08, Issue 01, Pages: 201–212.

Paper Authors

S. RADHA , G. SUNIL KUMAR

Vaishnavi Institute of Technology, Tirupati, AP, India.

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

INFORMATION RETRIEVAL FOR MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS

[1]S. RADHA , [2]G. SUNIL KUMAR
Dept. of CSE, Vaishnavi Institute of Technology, Tirupati, AP, India.
Email: radhasanakuppam@gmail.com

Abstract— In any aggressive business, achievement is primarily based at the capacity to make an object more attractive to clients than the competition. A wide variety of questions get up within the context of this undertaking: how do we formalize and quantify the competitiveness between objects? Who are the main competitors of a given object? What are the functions of an item that most have an effect on its competitiveness? Despite the effect and relevance of this trouble to many domains, best a constrained quantity of work has been committed toward an powerful answer. In this paper, we present a proper definition of the competitiveness between objects, primarily based available on the market segments that they could each cowl. Our evaluation of competitiveness makes use of purchaser evaluations, an considerable source of facts this is available in a huge variety of domains. We gift green methods for comparing competitiveness in massive evaluate datasets and address the herbal hassle of locating the pinnacle-ok competitors of a given object. Finally, we compare the fine of our outcomes and the scalability of our technique using a couple of datasets from special domains.

Index Terms—Data mining, Web mining, Information Search and Retrieval, Electronic commerce

1 INTRODUCTION

A Long line of research has validated the strategic importance of identifying and tracking a firm's competition [1]. Motivated by way of this problem, the advertising and marketing and control network have focused on empirical strategies for competitor identification [2], [3], [4], [5], [6], as well as on strategies for studying recognized competition [7].Extant research on the former has centered on mining comparative expressions (e.G. "Item A is better than Item B") from the Web or other textual assets [5], [6], [7], [1], [2], [3]. Even though such expressions can indeed be indicators of competitiveness, they're absent in lots of domains. For instance, don't forget the domain of vacation applications (e.G flight-hotel-automobile mixtures). In this case, gadgets haven't any assigned call by which they can be queried or compared with every other. Further, the frequency of textual comparative proof can vary significantly throughout domains. For example, while comparing logo names at the company stage (e.G. "Google vs Yahoo" or "Sony vs Panasonic"), it's miles certainly in all likelihood thatcomparative patterns can be observed with the aid of sincerely querying the net. However, it is easy to discover mainstream domain names wherein such evidence is extremely scarce, including footwear, jewelery, inns, eating places, and

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

fixtures. Motivated by way of these shortcomings, we endorse a new formalization of the competitiveness among two items, primarily based on the market segments that they could both cover. Formally:

Definition 1. **[Competitiveness]:** Let *U* be the population of all possible customers in a given market. We consider that an item *i* covers a customer *u 2 U* if it can cover all of the customer's requirements. Then, the competitiveness between two items *i, j* is proportional to the number of customers that they can both cover. Our competitiveness paradigm is based on the following

observation: The competitiveness among objects is based totally on whether or not they compete for the eye and enterprise of the identical corporations of customers (i.e. The equal market segments). For instance, two restaurants that exist in extraordinary countries are obviously no longer competitive, when you consider that there's no overlap among their target businesses. Consider the example shown in Figure 1.

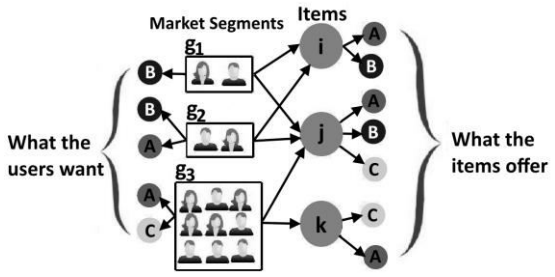

Fig. 1: A (simplified) example of our competitiveness paradigm

The figure illustrates the competitiveness between three items *i, j* and *k*. Each item is mapped to the set of features that it can offer to a customer. Three features are considered in this example: *A,B* and *C*. Even though this simple example considers only binary features (i.e. available/not available), our actual formalization accounts for a much richer space including binary, categorical and numerical features.

The left side of the figure shows three groups of customers *g*1, *g*2, and *g*3. Each group represents a different market segment. Users are grouped based on their preferences with respect to the features. For example, the customers in *g*2 are only interested in features *A* and *B*. We observe that items *i* and *k* are not competitive, since they simply do not appeal to the same groups of customers. On the other hand, *j* competes with both *i* (for groups *g*1 and *g*2) and *k* (for *g*3). Finally, an interesting observation is that *j* competes for 4 users with *i* and for 9 users with *k*. In other words, *k* is a stronger competitor for *j*, since it claims a much larger portion of its market share than *i*.This example illustrates the ideal scenario, in which we have access to the complete set of customers in a given market, as well as to specific market segments and their requirements. In practice, however, such information is not available. In order to overcome this, we describe a method for computing all the segments in a given market based on mining large review datasets. This method allows us to operationalize our definition of competitiveness and address the problem of finding the top-k competitors of an item in any given market. As we show in our work, this problem presents significant computational challenges, especially in the presence of large datasets with hundreds or thousands of items, such as those that are often found in mainstream domains. We address these challenges via a highly scalable framework for top-k computation,

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

including an efficient evaluation algorithm and an appropriate index.

Our work makes the following contributions:

• A formal definition of the competitiveness between two items, based on their appeal to the various customer segments in their market. Our approach overcomes the reliance of previous work on scarce comparative evidence mined from text.

• A formal methodology for the identification of the
different types of customers in a given market, as
well as for the estimation of the percentage of customers that belong to each type.

• A highly scalable framework for finding the top-k
competitors of a given item in very large datasets.

2 DEFINING COMPETITIVENESS

The typical user session on a review platform, such as Yelp, Amazon or Trip Advisor, consists of the following steps:

1) Specify all required features in a query.

2) Submit the query to the website's search engine and retrieve the matching items.

3) Process the reviews of the returned items and make a purchase decision.

In this setting, items that cover the user's requirements will be included in the search engine's response and will compete for her attention. On the other hand, non-covering items will not be considered by the user and, thus, will not have a chance to compete. Next, we present an example that extends this decision-making process to a multi-user setting. Consider a simple market with 3 hotels *i, j, k* and 6binary features: *bar, breakfast, gym, parking, pool, wi-fi*. Table 1 includes the value of each hotel for each feature. In this simple example, we assume that the market includes 6 mutually exclusive customer segments (types). Each segment is represented by a query that includes the features that are of interest to the customers included in the segment.

Information on each segment is provided in Table 2. For instance, the first segment includes 100 customers who are interested in parking and wi-fi, while the second segment includes 50 customers who are only interested in parking.

TABLE 1: Hotels and their Features.

Name	Bar	Breakfast	Gym	Parking	Pool	Wi Fi
Hilton	Yes	No	Yes	Yes	Yes	Yes
Blis	Yes	Yes	No	Yes	No	Yes
Grandridge	Yes	Yes	Yes	Yes	Yes	Yes

TABLE 2: Customer Segments

ID	Segment Size	Features of Interest
Q1	100	(*parking, wi-fi*)
Q2	50	(*parking*)
Q3	60	(*wi-fi*)
Q4	120	(*gym, wi-fi*)
Q5	250	(*breakfast, parking*)
Q6	80	(*gym, bar, breakfast*)

In order to measure the competition between any two hotels, we need to identify the number of customers that they can both satisfy. The results are shown in Table 3. The *Hilton* and the *Blis* can cover segments *q*1, *q*3, and *q*4. Therefore, they compete for (100 + 50 + 60)/660 _ 32% of the entire

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

market. We observe that this is the lowest competitiveness achieved for any pair, even though the two hotels are also the most similar. In fact, the highest competitiveness is observed between the *Bllis* and the *Grandridge*, that compete for 70% of the market. This is a critical observation that demonstrates that similarity is not a good proxy for competitiveness. The explanation is intuitive. The availability of both a pool and a bar makes the *Hilton* and the *Blis* more similar to each other and less similar to the *Grandrigde*. However, neither of these features has an effect on competitiveness. First, the *pool* feature is not required by any of the customers in this market. Second, even though the availability of a bar is required by segment *q*6, none of the three hotels can cover all three of this segment's requirements. Therefore, none of the hotels compete for this particular segment.Another intuitive observation is that the size of the segment has a direct effect on competitiveness. For example, even though the*Grandridge* shares the same number of segments (4) with the other two hotels, its competitiveness with the *Bllis* is significantly higher. This is due to the size of the *q*5 segment, which is more than double the size of *q*4.

TABLE 3: Common segments for restaurant pairs

Restaurant Pairs	**Common Segments**	**Common %**
Hilton, Blis	(q1,q2,q3)	32%
Hilton, Grandridge	(q1,q2,q3,q4)	50%
Blis,Grandridge	((q1,q2,q3,q5)	70%

The above example is limited to binary features. In this simple setting, it is trivial to determine if two items canboth cover a feature. However, as we discuss in detail in Section 2.1, the items in a market can have different types of features (e.g. numeric) that may be only *partially* covered by two items. Formally, let $p(q)$ be the percentage of users represented by a query q and let $V(i,j)\ q$ be the *pairwise coverage* offered by two items i and j to the space defined by the features in q. Then, we define the competitiveness between$i$ and j in a market with a feature subset F as follows:

$$(CF(i,j) \quad = \sum_{Q\in 2f}^{n} \mathrm{P(q)X\ V(i,j)Q} \quad (1)$$

This definition has a clear probabilistic interpretation:*given two items i, j, their competitiveness CF*(*i, j*) *represents the probability that the two items are included in the consideration set of a random user.* This new definition has direct implications for consumers, who often rely on recommendation systems to help them choose one of several candidate products. The ability to measure the competitiveness between two items enables the recommendation system to strategically select the order in which items should be recommended or the sets of items that should be included together in a group recommendation. For instance, if a random user *u* shows interest in an item *i*, then she is also likely to be interested in the items with the highest *CF*(*i, j*) values. Such competitive items are likely to meet the criteria satisfied by *i* and even cover additional parts of the feature space. In addition, as the user *u* rates more items and the system gains a more accurate view of her requirements, our

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

competitiveness measure can be trivially adjusted to consider only those features from F (and only those value intervals within each feature) that are relevant for u. This competitiveness-based recommendation paradigm is a departure from the standard approach that adjusts the weight (relevance) of an item j for a user u based on the rating that u submits for items
similar to j. As discussed, this approach ignores that
(i) the similarity may be due to irrelevant or trivial features and
(ii) for a user who likes an item i, an item j that is far superior than i with respect to the user's requirements (and thus quite different) is a better recommendation candidate than an item j' that is highly similar to i. In the following two sections we describe the computation of the two primary components of competitiveness: (1) the *pairwise coverage* $V\ q\ (i,j\ )$of a query that includes binary, categorical, ordinal or numeric features, and (2) the percentage $p(q)$ of users represented by each query q.

2.1 Pairwise Coverage

We begin by defining the pairwise coverage of a single feature f. We then define the pairwise coverage of an entire *query* of features q.

Definition 2. **[Pairwise Feature Coverage]**: We define the *pairwise coverage* $V\ f\ (i,j\ )$of a feature f by two items i, j as the percentage of all possible values of f that can be covered by *both* i and j. Formally, given the set of all possible values $V\ f$ for f, we define:

$$V_{i,j}^{f} = |\{V \quad V \in Vf : v\mathcal{L}f(i) \Lambda v\mathcal{L}f[j]\}/ values(f)$$

where $v\mathcal{L}f[i]$ represents that v is *covered* by the value of item i for feature f.

[Binary and Categorical Features]: Categorical features take one or more values from a finite space. Examples of single value features include the brand of a digital camera or the location of a restaurant. Examples of multi-value features include the amenities offered by a hotel or the types of cuisine offered by a restaurant. Any categorical feature can be encoded via a set of binary features, with each binary feature indicating the (lack of) coverage of one of the original feature's possible values. In this simple setting, the feature can be fully covered (if $f[i] = f[j] = 1$ or, equivalently, $f[i]\_f[j] = 1$), or not covered at all. Formally, the pairwise coverage of a binary feature f by two items i, j can be computed as follows:

$$V\,fi,j = \min(f[i], f[j]) \text{(binary features)} \quad (2)$$

[Numeric Features]: Numeric features take values from a pre-defined range. Henceforth, without loss of generality, we consider numeric features that take values in [0, 1], with higher values being preferable. The pairwise coverage of a numeric feature f by two items i and j can be easily computed as the smallest (worst) value achieved for f by either item. For instance, consider two restaurants i, j with values 0.8 and 0.5 for the feature *food quality*. Their pairwise coverage in this setting is 0.5. Conceptually, the two items will compete for any customer who accepts a quality ≤0.5. Customers with higher standards would eliminate restaurant j, which will never have a chance to compete for their business. Formally, the pairwise coverage of a numeric

$$Vi,jf = \min(f(i), f(j)) \quad (3)$$

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

feature *f* by two items *i, j* can be computed as follows: $V_{fi,j} = \min(f[i], f[j])$ (numeric features) (3)

[Ordinal Features]: Ordinal features take values from a finite *ordered* list. A characteristic example is the popular five star scale used to evaluate the quality of a service or product. For example, consider that the values of two items *i* and *j* on the 5-star rating scale are ⋆⋆ and ⋆⋆⋆, respectively. Customers that demand at least 4 stars will not consider either of the two items, while customers that demand at least 3 stars will only consider item *j*. The two items will thus compete for all customers that are willing to accept 1 or 2 stars. Therefore, as in the case of numeric features, the pairwise coverage for ordinal features is determined by the worst of the two values. In this example, given that the two items compete for 2 of the 5 levels of the ordinal scale (1 and 2 stars), their competitiveness is proportional to 2⁄5 = 0.4. Formally, the pairwise coverage of an ordinal feature *f* by two items *i, j* can be computed as follows:

Vi,jf=min(f(i),f(j))/|Vf)|

(4)

Pairwise coverage of a feature query: We now discuss how coverage can be extended to the query level. Figure 2 visualizes a query *q* that includes two numeric features *f*1 and *f*2. The figure also includes two competitive items *i* and *j*, positioned according to their values for the two features:

*f*1[*i*] = 0.3, *f*2[*i*] = 0.3, *f*1[*j*] = 0.2, and *f*2[*j*] = 0.7.

We observe that the percentage of the 2-dimensional space that each item covers is equivalent to the area of the rectangle defined by the beginning of the two axes (0, 0) and the item's values for *f*1 and *f*2. For example, the covered area for item *i* is 0.3_0.3 = 0.09, equal to 9% of the entire space. Similarly, the pairwise coverage provided by both items is equal to 0.2 _ 0.3 = 0.06 (i.e. 6% of the market). Per our example, the pairwise coverage of a given query *q* by two items *i, j* can be measured as the volume of the

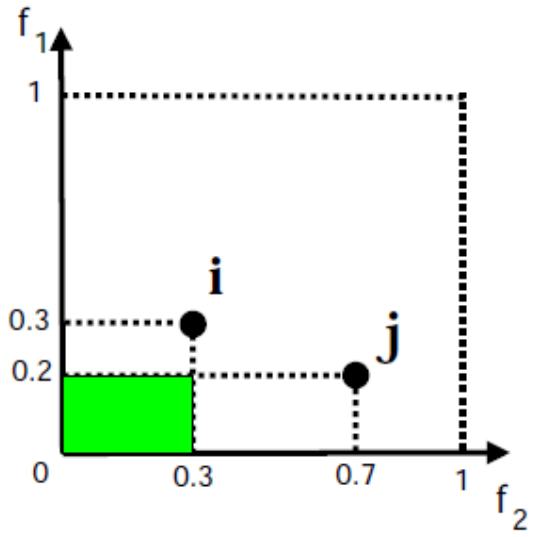

Fig. 2: Geometric interpretation of pairwise coverage

hyper-rectangle defined by the pairwise coverage provided by the two items for each feature *f* 2€ *q*. Formally:Empirical Model And Measurement Variables In response to the gap in extant research, we prose the following empirical model as a way to conceptualize firm-side authenticity across contexts and environments. Drawing on well-established characterizations of the construct, our framework is based on the Competitive Advantage (CA) that results from congruence between the internal, withinfirm, strategically-directed aspects of an organization's culture, orientation, and

use of resources and capabilities; operationalized as Innovation Capacity (IC)—i.e., Beverland's (2005) Partly true, alongside the outwardly-directed branding-building and marketing communications activities operationalized as Corporate Identity Management (CIM).— i.e., Beverland's (2005) Partly rhetorical. Further rationale for our choice of each variable is discussed in greater detail below. 9 Innovation Capacity (IC) Hurley and Hult (1998), and other authors (e.g., Kogut and Zander, 1992; Kov and Ceylan, 2007; Calantone, Cavusgil, and Zhao, 2002) employ the construct of Innovation Capacity (IC) to capture the broad range of within-firm characteristics that contribute to a firm's capacity to innovate. Accordingly, IC measures the firm's orientation, culture, and ability to mobilize and combine resources, capabilities, and knowledge to generate new ideas-- such as creating new products, services, or process innovations (e.g., Lawson and Samson, 2001; Hurley and Hult, 1998; Martensen and Dahlgaard, 1999). We follow related research who highlight how IC is a reliable measure of how organizations are able to build and maintain consistent cultures through the combination of employee skills and expertise with firm resources, competencies, and capabilities as directed toward strategic objectives (Hauser, 1998; Prajogo and Ahmed, 2006; Chen, Lin, and Chang, 2009). Accordingly, IC has been demonstrated to be a major influence on firm performance (Herbig and Dunphy, 1998; Deshpande and Gatingon, 1994; Parkman, Holloway and Sebastio, 2012; Johnson, 2001; McEvily and Chakravarthy, 2002; Berghman, Matthyssens, Streukens, and Vandenbempt, 2013). We employ this construct as a proxy to capture both the general desire of organizations (as supported by firm culture, orientation, and philosophy) and the specific strategic decisions made by firm managers to direct allocations of resources towards product and service innovations that support the firm authentically Being true to itself (in contrast to those offerings which are "fake", "copycat", or "phony").

RESEARCH METHOD Research Context Our research setting is architectural design services firms (NAICS code: 541310). For several reasons we contend that architectural design appears well suited for our firm-level examination of IC and CIM, as well as the strategic use of firm-side authenticity. First, architecture provides an environment where organizational-level use of creativity, innovation, and firm-side authenticity are central strategic imperatives (Howkins, 2002; Caves, 2003; Gilson and Shalley, 2004). Firms in architectural design services vary in their organizational cultures and orientations as well as in their creative and innovative competencies. Second, the architecture design services firms in our sample are profitdriven and contest highly competitive marketplaces. The implications of these factors are that firms have clear incentives to direct scarce resources towards the most commercially viable offerings-- in contrast to non-profit creative or artistic organizations (O'Reilly, Rentschler, and Kirchner, 2013; Fillis, 2003; Fillis and Rentschler, 2006; Boyle, 2007). Third, because organizations possess differentiated resources and capabilities and the marketplace is highly competitive, firm

performance and competitive advantage has been shown to be influenced by beneficial corporate images and reputation, brand building and marketing promotions (Caves, 2002; Bilton and Leary, 2004). Taken together, architectural design services provides a useful context for our exploratory study; where firms and their offerings compete against one another largely on the basis of co-created 13 meaning between an object and consumer helping to distinguish real from fake (Peterson 2005) —i.e., authenticity.

3 Datasets and Baselines

Our experiments include four datasets, which were collected for the purposes of this project. The datasets were intentionally selected from different domains to portray the cross-domain applicability of our approach. In addition to the full information on each item in our datasets, we also collected the full set of reviews that were available on the source website. These reviews were used to (1) estimate queries probabilities, as described in Section 2.2 and (2) extract the opinions of reviewers on specific features.The highly-cited method by Ding et al. [28] is used to convert each review to a vector of opinions, where each opinion is defined as a feature-polarity combination (e.g. service+, food-). The percentage of reviews on an item that express a positive opinion on a specific feature is used as the feature's numeric value for that item. We refer to these as *opinion features*. Table 4 includes descriptive statistics for each dataset, while a detailed description is provided below.

CAMERAS: This dataset includes 579 digital cameras from Amazon.com. We collected the full set of reviews for each camera, for a total of 147192 reviews. The set of features includes the *resolution* (in MP), *shutter speed* (in seconds), *zoom* (e.g. 4x), and *price*. It also includes opinion features on *manual, photos, video, design, flash, focus, menu options, lcdscreen, size, features, lens, warranty, colors, stabilization, batterylife, resolution*, and *cost*.

HOTELS: This dataset includes 80799 reviews on 1283 hotels from Booking.com. The set of features includes the *facilities*, *activities*, and *services* offered by the hotel. All three of these multi-categorical features are available on the website.

The dataset also includes opinion features on *location, services, cleanliness, staff*, and *comfort*.

RESTAURANTS: This dataset includes 30821 reviews on 4622 New York City restaurants from TripAdvisor.com. The set of features for this dataset includes the *cuisine types* and *meal types* (e.g. lunch, dinner) offered by the restaurant, as well as the *activity types* (e.g. drinks, parties) that it is good for. All three of these multi-categorical features are available on the website. The dataset also includes opinion features on *food,service, value-for-money, atmosphere*, and *price*.

RECIPES: This dataset includes 100000 recipes from Sparkrecipes.com. It also includes the full set of reviews on each recipe, for a total of 21685 reviews. The set of features for each recipe includes the *number of calories*, as well as the following nutritional information, measured in grams: *fat, cholesterol, sodium, potassium, carb, fiber, protein, vitamin A,*

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

vitamin B12, vitamin C, vitamin E, calcium, copper, folate,magnesium, niasin, phosphorus, riboflavin, selenium, thiamin, zinc. All information is openly available on the website.

TABLE 4: Dataset Statistics

dataset	Items	Feast	Skyline Layers	Subsets
CAMERAS	520	20	5	14700
HOTELS	25	8	5	129
RESTAURANTS	100	3	12	158
RECIPES	10000	5	25	2000

For each dataset, the 2nd, 3rd, 4th and 5th columns include the number of items, the number of features, the number of distinct queries, and the number of layers in the respective skyline pyramid, respectively. In order to conclude the description of our datasets, we present some statistics on the skyline-pyramid structure constructed for each corpus. Figure 4 shows the distribution of items in the first 6 skyline layers of each dataset. We observe that, for

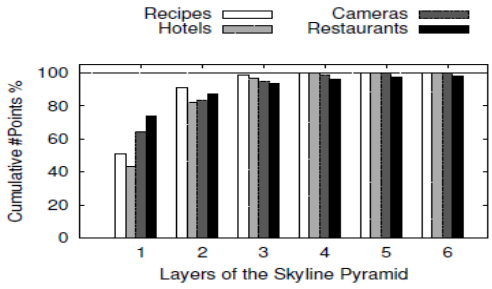

Fig. 3: Cumulative distribution of items across the first 6 layers of the skyline pyramid.

all datasets, nearly 99% of the items can be found within the first 4 layers, with the majority of those falling within the first 2 layers. This is due to the large dimensionality of the feature space, which makes it difficult for items to dominate one another. As we show in our experiments, the skyline pyramid enables CMiner to clearly outperform the

baselines with respect to computational cost. This is despite the high concentration of items within the first layers, since CMiner can effectively traverse the pyramid and consider only a small fraction of these items.

Baselines: We compare CMiner with two baselines. The *Naive* basline, is the brute-force approach described in Section 3. The second is a clustering-based approach that first iterates over every query q and identifies the set of items that have the same value assignment for the features in q and places them in the same group. The algorithm then iteratesover the reported groups and updates the pairwise coverage $Vq\ i,j$ for the item of interest i and an arbitrary item j from each group (it can be any item, since they all have the same values with respect to q). The computed coverage is then used to update the competitiveness of all the items in the group. The process continues until the final competitiveness scores for all items have been computed. Assuming that we have a collection of items I, a set of queries Q, and at most M groups per query, the complexity is O($jIj * M * jQj$)

4 Computational Time

In this experiment we compare the speed of CMiner with that of the two baselines (Naive and GMiner), as well as with that of

the enhanced CMiner++ algorithm. Specifically, we use each algorithm to compute the set of top-k competitors for each item in our datasets. The results are shown in Figure. 5. Each plot reports the average time, in seconds, per item (y-axis) against the various *k* values (x-axis).The figures motivate some interesting observations. First, the Naive algorithm consistently reports the same computational time regardless of *k*, since it naively computes the competitiveness of every single item in the corpus with respect to the target item. Thus, any trivial variations in the required time are due to the process of maintaining the top-*k* set. In general, Naive is outperformed by the two other algorithms, and is only competitive for very large values of *k* for the HOTELS dataset. The latter case can be attributed to the small number of queries and items included in this dataset, which limit the ability of more sophisticated algorithms to significantly prune the space when the number of required competitors is very large. For the CAMERAS dataset, CMiner and GMiner, exhibit almost identical running times. This is due to (1) the very large number of distinct queries for this dataset (14779), which serves as a computational bottleneck for CMiner and (2) the highly clustered structure of the item population, which includes 579 items. A deeper analysis reveals that GMiner identifies and average of 443.63 item groups (i.e. groups of identical items) per query. This means that the algorithm saves (on expectation) a total of (579x443)-14779 = 241718 coverage computations per query, allowing it to be competitive to the otherwise superior CMiner. In fact, for the other datasets, CMiner displays a clear advantage. This advantage is maximized for the RECIPES dataset, which is the most populous of the four in terms of included items. The experiment on this dataset also illustrates the scalability of the approach with respect to *k*. For the HOTELS and RESTAURANTS datasets, even though the computational time of CMiner appears to rise as *k* increases for the other three datasets, it never goes above 0.035 seconds. For the CAMERAS dataset, the large number of considered queries has an adverse of the scalability of CMiner, since it results in larger number of required computations for larger values of *k*. This finding motivates us to consider pruning the set of queries by eliminating those that have a low probability. We explore this direction in the experiment presented in Section 5.6. Finally, we observe that the enhanced CMiner++ algorithm consistently outperformed all the other approaches, across datasets and values of *k*. The advantage of CMiner++ is increased for larger values of *k*, which allow the algorithm to benefit from its improved pruning. This verifies the utility of the improvements described in Section 4.2 and demonstrates that effective pruning can lead to a performance that far exceeds the worst-case complexity analysis of CMiner.

5. Computational Time

In this experiment we compare the speed of CMiner with that of the two baselines (Naive and GMiner), as well as with that of the enhanced CMiner++ algorithm. Specifically,we use each algorithm to compute the set of top-k competitors for each item in our datasets. The results are shown in Figure. 5. Each plot reports the

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

average time, in seconds, per item (y-axis) against the various k values (x-axis). The figures motivate some interesting observations. First, the Naive algorithm consistently reports the same computational time regardless of k, since it naively computes the competitiveness of every single item in the corpus with respect to the target item. Thus, any trivial variations in the required time are due to the process of maintaining the top-k set. In general, Naive is outperformed by the two other algorithms, and is only competitive for very large values of k for the HOTELS dataset. The latter case can be attributed to the small number of queries and items included in this dataset, which limit the ability of more sophisticated algorithms to significantly prune the space when the number of required competitors is very large. For the CAMERAS dataset, CMiner and GMiner, exhibit almost identical running times. This is due to (1) the very large number of distinct queries for this dataset (14779), which serves as a computational bottleneck for CMiner and (2) the highly clustered structure of the item population, which includes 579 items. A deeper analysis reveals that GMiner identifies and average of 443.63 item groups (i.e. groups of identical items) per query. This means that the algorithm saves (on expectation) a total of$(579 \; x \; 443) _ 14779 = 2009944$ coverage computations per query, allowing it to be competitive to the otherwise superior CMiner. In fact, for the other datasets, CMiner displays a clear advantage. This advantage is maximized for the RECIPES dataset, which is the most populous of the four in terms of included items. The experiment on this dataset also illustrates the scalability of the approach with respect to k. For the HOTELS and RESTAURANTS datasets, even though the computational time of CMiner appears to rise as k increases for the other three datasets, it never goes above 0.035 seconds. For the CAMERAS dataset, the large number of considered queries has an adverse of the scalability of CMiner, since it results in larger number of required computations for larger values of k. This finding motivates us to consider pruning the set of queries by eliminating those that have a low probability. We explore this direction in the experiment presented in Section 5.6. Finally, we observe that the enhanced CMiner++ algorithm consistently outperformed all the other approaches, across datasets and values of k. The advantage of CMiner++ is increased for larger values of k, which allow the algorithm to benefit from its improved pruning. This verifies the utility of the improvements described in Section 4.2 and demonstrates that effective pruning can lead to a performance that far exceeds the worst-case complexity analysis of CMiner.

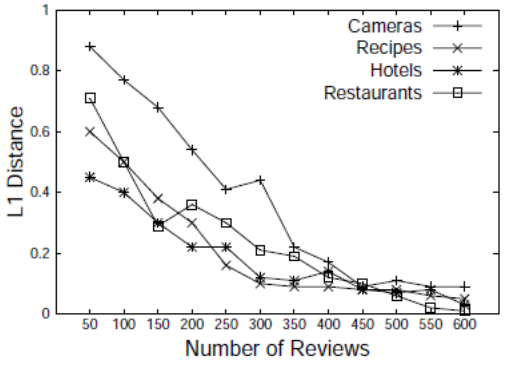

Fig. 4: Convergence of query probabilities.

Based on our results, we see that all the datasets exhibited near identical trends. This is an encouraging finding with useful implications, as it informs us that any

conclusions we draw about the convergence of the computed probabilities will be applicable across domains. Second, the figures clearly demonstrate the convergence of the computed probabilities, with the reported L1 distance dropping rapidly to trivial levels below 0.2, after the consideration of less than 500 reviews. The convergence of the probabilities is an especially encouraging outcome that (i) reveals a stable categorical distribution for the preferences of the users over the various queries, and (ii) demonstrates that only a small seed of reviews, that is orders of magnitude smaller than the thousands of reviews available in each dataset, is sufficient to achieve an accurate estimation of the probabilities.

6 CONCLUSION

We presented a formal definition of competitiveness between two items, which we validated both quantitatively and qualitatively. Our formalization is applicable across domains, overcoming the shortcomings of previous approaches. We consider a number of factors that have been largely overlooked in the past, such as the position of the items in the multi-dimensional feature space and the preferences and opinions of the users. Our work introduces an end-to-end methodology for mining such information from large datasets of customer reviews. Based on our competitiveness definition, we addressed the computationally challenging problem of finding the top-k competitors of a given item. The proposed framework is efficient and applicable to domains with very large populations of items. The efficiency of our methodology was verified via an experimental evaluation on real datasets from different domains. Our experiments also revealed that only a small number of reviews is sufficient to confidently estimate the different types of users in a given market, as well the number of users that belong to each type.

REFERENCE

1. Beer, S. (2008). Authenticity and food xperience–commercial and academic perspectives. Journal of Foodservice, 19(3), 153-163.
2. R. Deshpand and H. Gatingon, “Competitive analysis,” *Marketing Letters*, 1994.
3. M. Bergen and M. A. Peteraf, “Competitor identification and competitor analysis: a broad-based managerial approach,” *Managerial and Decision Economics*, 2002.
4. J. F. Porac and H. Thomas, “Taxonomic mental models in competitor definition,” *The Academy of Management Review*, 2008.
5. M.-J. Chen, “Competitor analysis and interfirm rivalry: Toward a theoretical integration,” *Academy of Management Review*, 1996.
6. Balmer, J. M. T. (2001). Corporate identity, corporate branding and corporate marketing: seeing through the fog, European Journal of Marketing, 35, pp. 248–292.