MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS

SHANMUKHI SAI NAINAR[1]

YARRAGUNTLA SAISWETHA[2]

[1]BTech Student, RMK College of Engineering, RSM nagar, Gummadipoondi Taluk, Tiruvallur district, Kavaraipettai, Tamilnadu 601206, India.

[2]BTech Student, R V R & J C College of Engineering, Chandramoulipuram, Chowdavaram, Guntur-522 019, Andhra Pradesh, India

Abstract: In any type of affordable company, success is based upon the capacity to make a product extra enticing to clients than the competitors. A variety of concerns emerge in the context of this job: exactly how do we define and also evaluate the competition in between 2 things? That are the major rivals of a provided product? What are the functions of a product that many impact its competition? Regardless of the influence as well as importance of this trouble to several domain names, just a restricted quantity of job has actually been dedicated towards a reliable remedy. In this paper, we provide an official interpretation of the competition in between 2 things, based upon the marketplace sections that they can both cover. Our examination of competition uses consumer testimonials, a mother lode of info that is offered in a vast array of domain names. We offer reliable approaches for assessing competition in big evaluation datasets and also deal with the all-natural issue of discovering the top-k rivals of a provided product. Lastly, we examine the top quality of our outcomes as well as the scalability of our strategy making use of numerous datasets from various domain names.

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

I. INTRODUCTION

Along line of research study has actually shown the tactical relevance of determining and also keeping an eye on a company's rivals [1] Encouraged by this issue, the advertising as well as administration area have actually concentrated on empirical approaches for rival recognition [2], [3], [4], [5], [6], in addition to on approaches for examining recognized rivals [7] Extant research study on the previous has actually

concentrated on mining relative expressions (e.g. "Product A is much better than Product B") from the Internet or various other textual resources. Despite the fact that such expressions can undoubtedly be indications of competition, they are missing in several domain names. For example, take into consideration the domain name of trip plans (e.g flight-hotel-car mixes). In this instance, products have actually no appointed name through which they can be inquired or compared to each various other. Better, the regularity of textual relative proof can differ considerably throughout domain names. As an example, when contrasting brand at the company degree (e.g. "Google vs Yahoo" or "Sony vs Panasonic"), it is without a doubt most likely that relative patterns can be located by merely inquiring the internet. Nonetheless, it is very easy to recognize mainstream domain names where such proof is exceptionally limited, such as footwear, jewelery, resorts, dining establishments, as well as furnishings. Encouraged by these imperfections, we recommend a brand-new formalization of the competition in between 2 things, based upon the marketplace sectors that they can both cover. Officially:

Competitiveness: Allow U be the populace of all feasible clients in a provided market. We think about that a thing i covers a consumer u 2 U if it can cover every one of the client's demands. After that, the competition in between 2 things i, j is symmetrical to the variety of clients that they can both cover.

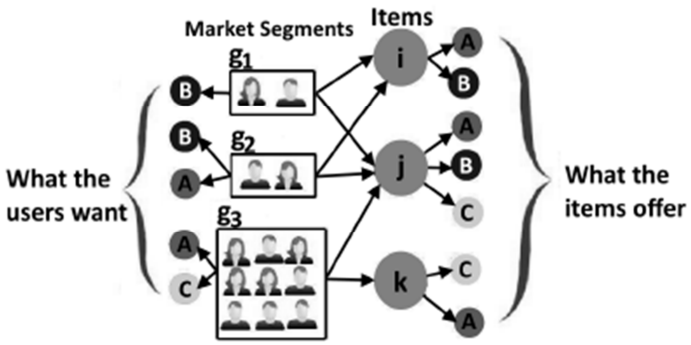

Fig 1: A (simplified) example of our competitiveness paradigm

Our competition standard is based upon the complying with monitoring: the competition in between 2 things is based upon whether they complete for the interest as well as service of the very same teams of clients (i.e. the exact same market sectors). As an example, 2 dining establishments that exist in various nations are clearly not affordable, considering that there is no overlap in between their target teams. Take into consideration the instance received Number 1. The number highlights the competition in between 3 products i, j as well as k. Each product is mapped to the collection of functions that it can use to a consumer. 3 functions are thought about in this instance: A, B as well as C. Despite the fact that this

straightforward instance takes into consideration just binary attributes (i.e. available/not readily available), our real formalization make up a much richer area consisting of binary, specific and also mathematical attributes. The left side of the number reveals 3 teams of consumer's g1, g2, as well as g3. Each team stands for a various market sector. Customers are organized based upon their choices relative to the attributes. As an example, the consumers in g2 are just thinking about attributes An and also B. We observe that things i and also k are not affordable, given that they just do not interest the very same teams of consumers. On the various other hand, j takes on both i (for teams g1 as well as g2) as well as k (for g3). Lastly, a fascinating monitoring is that j contends for 4 individuals with i as well as for 9 individuals with k. To put it simply, k is a more powerful rival for j, because it asserts a much bigger section of its market share than i. This instance shows the optimal circumstance, in which we have accessibility to the total collection of clients in a provided market, along with to details market sectors as well as their demands. In technique, nevertheless, such details is not offered. In order to conquer this, we define a technique for calculating all the sectors in an offered market based upon mining big evaluation datasets. This approach enables us to operationalize our meaning of competition as well as attend to the trouble of locating the top-k rivals of a product in any type of provided market. As we display in our job, this issue offers considerable computational obstacles, specifically in the visibility of huge datasets with hundreds or countless things, such as those that are frequently discovered in mainstream domain names. We deal with these obstacles through an extremely scalable structure for top-k calculation, consisting of an effective examination formula and also a suitable index.

II. RELATED WORK

This paper improves as well as dramatically prolongs our initial work with the analysis of competition. To the most effective of our expertise, our job is the very first to attend to the analysis of competition using the evaluation of huge disorganized datasets, without the requirement for straight relative proof. However, our job has connections to previous job from numerous domain names.

Managerial Competitor Identification: The administration literary works is abundant with jobs that concentrate on exactly how supervisors can by hand

determine rivals. Several of these jobs version rival recognition as a psychological classification procedure in which supervisors developing depictions of rivals as well as utilize them to identify prospect companies. Various other hands-on classification approaches are based upon market- and also resource-based resemblances in between a company as well as prospect rivals [1], [5], [7] Lastly, supervisory rival recognition has actually likewise existed as a sense making procedure in which rivals are recognized based upon their possible to intimidate a company's identification [4]

Competitor Mining Algorithms: Zheng et al recognize crucial affordable procedures (e.g. market share, share of pocketbook) as well as demonstrated how a company can presume the worth's of these steps for its rivals by mining (i) its very own comprehensive consumer deal information and also (ii) accumulated information for each and every rival. Unlike our very own technique, this method is not ideal for reviewing the competition in between any type of 2 things or companies in a provided market. Rather, the writers presume that the collection of rivals is offered as well as, therefore, their objective is to calculate the worth of the picked procedures for each and every rival. Furthermore, the reliance on transactional information is a restriction we do not have. Doan et al. discover individual visitation information, such as the geo-coded information from location-based social media networks, as a prospective source for rival mining. While they report encouraging outcomes, the dependancy on visitation information restricts the collection of domain names that can gain from this strategy. Pant and also Sheng assume and also validate that contending companies are most likely to have comparable internet impacts, a sensation that they describe as online isomorphism Their research takes into consideration various sorts of isomorphism in between 2 companies, such as the overlap in between the in-links as well as out links of their corresponding web sites, along with the variety of times that they show up with each other on-line (e.g. in search engine result or brand-new posts). Comparable to our very own method, their technique is tailored towards pair wise competition. Nonetheless, the demand for isomorphism functions restricts its applicability to companies and also make it improper for things as well as domain names where such functions are either not offered or very sporadic, as is generally the situation

with co-occurrence information. Actually, the sparsity of co-occurrence information is a serious limitation of a considerable body of job that concentrates on mining rivals based upon relative expressions discovered in internet outcomes as well as various other textual corpora. The instinct is that the regularity of expressions like "Thing A is much better than Product B" "or product A Vs. Thing B" is a measure of their competition. Nevertheless, as we have actually currently talked about in the intro, such proof is usually limited and even non-existent in numerous traditional domain names. Therefore, the applicability of such strategies is substantially restricted. We give empirical proof on the sparsity of co-occurrence details in our speculative examination.

Finding Competitive Products: Current job has actually discovered competition in the context of item layout. The initial step in these methods is the interpretation of a prominence feature that stands for the worth of an item. The objective is after that to utilize this feature to develop products that are not controlled by various other, or optimize things with the optimum feasible supremacy worth. A comparable kind of work [39], [40] stands for products as factors in a multidimensional room as well as seeks subspaces where the charm of the thing is made the most of. While appropriate, the above jobs have an entirely various emphasis from our very own, and also thus the recommended techniques are not appropriate in our setup.

Skyline computation: Our job leverages principles as well as methods from the substantial literary works on horizon calculation these consist of the supremacy idea amongst things, in addition to the building and construction of the sky line pyramid utilized by our CMiner formula. Our job likewise has connections to the current magazines backwards sky line inquiries despite the fact that the emphasis of our job is various, we mean to make use of the breakthroughs in this area to enhance our structure in future job.

III. PROPOSED MODEL

COMPETITIVENESS: The normal customer session on a testimonial system, such as Yelp, Amazon.com or Journey Consultant, contains the complying with actions: Define all called for functions in a question. Send the inquiry to the web site's online search engine and also fetch the matching things. Refine the testimonials of the returned products and also purchase choice.

Pairwise Coverage

We start by specifying the pairwise protection of a solitary function f. We after that specify the pairwise insurance coverage of a whole question of attributes q.

Pairwise Function Protection: We specify the pairwise protection V f i, j of an attribute f by 2 things i, j as the portion of all feasible worths of f that can be covered by both i and also j. Officially, provided the collection of all feasible worths V f for f, we specify:

$$V_{i,j}^{f} = \frac{|\{v \in V^{f} : v \angle f[i] \wedge v \angle f[j]\}|}{|values(f)|},$$

where vf[i] represents that v is covered by the value of item i for feature f. Next, we describe the computation of V f i,j for different types of features.

Binary and Categorical Features: Specific functions take several worths from a limited room. Instances of singlevalue attributes consist of the brand name of an electronic cam or the area of a dining establishment. Instances of multi-value functions consist of the features supplied by a resort or the sorts of food used by a dining establishment. Any type of specific function can be inscribed by means of a collection of binary functions, with each binary function suggesting the (absence of) protection of among the initial function's feasible worths. In this straightforward setup, the attribute can be completely covered (if f [i] = f [j] = 1 or, equivalently, f [i] _ f [j] = 1), or otherwise covered whatsoever. Officially, the pairwise protection of a binary attribute f by 2 products i, j can be calculated as adheres to:

$$V_{i,j}^{f} = f[i] \times f[j] \quad \text{(binary features)} \qquad (2)$$

Numeric Features]: Numerical functions take worths from a. Pre-defined array. Henceforth, without loss of generalization, we think about numerical attributes that take worths in [0, 1], with greater worths being more effective. The pairwise protection of a numerical attribute f by 2 things i as well as j can be quickly calculated as the tiniest (worst) worth attained for f by either product. As an example, think about 2 dining establishments i, j with worths 0.8 and also 0.5 for the function food top quality. Their pairwise protection in this setup is 0.5. Conceptually, both products will certainly contend for any kind of consumer that approves a high quality _ 0.5. Clients with greater requirements would certainly remove dining establishment j, which will certainly never ever have an opportunity to complete for their company. Officially, the pairwise insurance coverage of a numerical attribute f by 2 products i, j can be calculated as adheres to:.

$$V_{i,j}^{f} = \min(f[i], f[j]) \quad \text{(numeric features)} \qquad (3)$$

[Ordinal Features]: Ordinal functions take worths from a limited bought listing. A particular instance is the prominent 5 star range made use of to review the top quality of a product or service. As an example, think about that the worths of 2 things i as well as j on the 5-star score range are ⋆ ⋆ as well as ⋆ ⋆ ⋆, specifically. Clients that require at the very least 4 celebrities will certainly rule out either of both things, while clients that need a minimum of 3 celebrities will just take into consideration thing j. Both things will certainly hence complete for all clients that want to approve 1 or 2 celebrities. For that reason, as when it comes to numerical attributes, the pairwise insurance coverage for ordinal attributes is figured out by the worst of both worths. In this instance, considered that both products contend for 2 of the 5 degrees of the ordinal range (1 and also 2 celebrities), their competition is symmetrical to 2/5 = 0.4. Officially, the pairwise protection of an ordinal function f by 2 products i, j can be calculated as complies with:

$$V_{i,j}^{f} = \frac{\min(f[i], f[j])}{|V^{f}|} \quad \text{(ordinal features)}$$

Pairwise coverage of a feature query: We currently go over just how protection can be encompassed the question degree. Number 2 envisions a question q that consists of 2 numerical attributes f1 and also f2. The number additionally consists of 2 affordable products i and also j, located according to their worths for both functions: f1 [i] = 0.3, f2 [i] = 0.3, f1 [j] = 0.2, and also f2 [j] = 0.7. We observe that the portion of the 2-dimensional room that each product covers amounts the location of the rectangular shape specified by the start of both axes (0, 0) and also the thing's worths for f1 and also f2. For instance, the protected location for product i is 0.3 _ 0.3 = 0.09, equivalent to 9% of the whole area. Likewise, the pairwise protection given by both products amounts to 0.2 _ 0.3 = 0.06 (i.e. 6% of the marketplace). Per our instance, the pairwise insurance coverage of a provided question q by 2 things i, j can be determined as the quantity of the

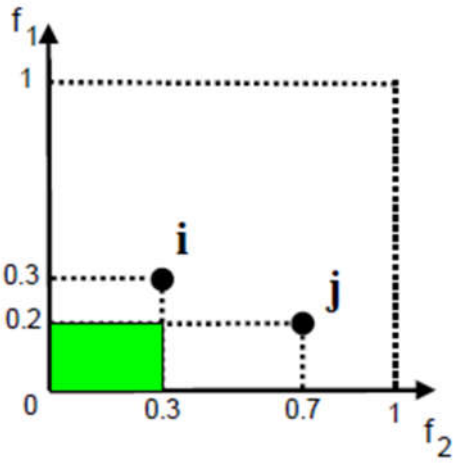

Fig. 2: Geometric interpretation of pairwise coverage

hyper-rectangle defined by the pairwise coverage provided by the two items for each feature f 2 q. Formally:

$$V_{i,j}^{q} = \prod_{f \in q} V_{i,j}^{f} \qquad (5)$$

Eq. 5 allows us to compute the pairwise coverage of any query of features, as required by the definition of competitiveness in Eq. 1.

Estimating Query Probabilities

The interpretation of competition given up Eq. 1 thinks about the possibility p(q) that an arbitrary consumer will certainly be stood for by a particular question of attributes q, for each feasible question q 2 2F. In this area, we explain just how these possibilities can be approximated from genuine information. Function questions are a straight depiction of individual choices. Preferably, we would certainly have accessibility to the inquiry logs of the system's (e.g. Amazon.com's or TripAdvisor's) online search engine. In technique, nonetheless, the delicate as well as exclusive nature of such info makes it extremely hard for companies to share openly. Consequently, we develop an evaluation procedure that just needs accessibility to a bountiful source: consumer testimonials. Each evaluation consists of a consumer's viewpoints on a certain part of attributes of the evaluated thing.

Extending our Competitiveness Definition

Feature Uniformity: Our competition meaning presumes that customer needs are consistently dispersed within the worth area of each function. This presumption enables us to develop a computational design for competition, yet in method it might not constantly hold true. As an example, the variety of customers requiring top quality in [0, 0.1] may be various than those requiring a worth in [0.4, 0.5] In addition, for absence of even more precise info, it supplies a traditional reduced bound of our design's real efficiency: having accessibility to the circulation of rate of interest within each attribute can just boost the top quality of our outcomes. If such info was undoubtedly offered, after that the naïve method would

certainly be to take into consideration all feasible passion periods mixes for all feasible questions. Henceforth, we describe these as extensive questions. Plainly, the variety of feasible extensive inquiries is rapid as well as provides the computational expense of any type of examination formula expensive.

FINDING THE TOP-K COMPETITORS: Given the definition of the competitiveness in Eq. 1, we study the natural problem of finding the top-k competitors of a given item. Formally:

Problem 1. [Top-k Competitors Problem]: We exist with a market with a collection of n products I as well as a collection of functions F. After that, provided a solitary thing i 2 I, we intend to recognize the k products from I that take full advantage of CF(i, _). An ignorant formula would certainly calculate the competition in between i and also every feasible prospect. The intricacy of this strength approach is clear, which can be quickly controlled by the powerset variable and also, as we show in our experiments, is unwise for big datasets. One choice might be to execute the naïve calculation in a dispersed style. Also in this instance, nonetheless, we would certainly require one string for each and every of the n2 sets. This is much from minor, if one thinks about that n might determine in the 10s of thousands. Furthermore, a naïve MapReduce execution would certainly deal with the traffic jam of passing whatever with the reducer to represent the self-join consisted of in the calculation. In technique, the selfjoin would certainly need to be carried out through a personalized method for reduce-side signs up with, which is a non-trivial and also extremely costly procedure [23] These problems encourage us to present CMiner, an effective specific formula for Issue 1. With the exception of the development of our indexing device, every various other element of CMiner can additionally be integrated in an identical remedy. Initially, we specify the principle of product supremacy, which will certainly assist us in our evaluation:

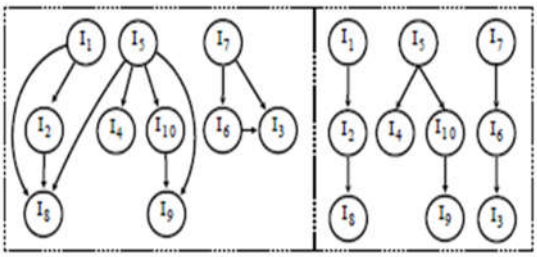

Fig 3: The left side shows the dominance graph for a set of items. An edge Ii→Ij means that Ii dominates Ij . The right side of the figure shows the skyline pyramid.

The CMiner Algorithm: Next off, we offer CMiner, a specific formula for discovering the top-k rivals of an offered thing. Our

formula utilizes the horizon pyramid in order to decrease the variety of things that require to be taken into consideration. Considered that we just appreciate the top-k rivals, we can incrementally calculate ball game of each prospect and also quit when it is ensured that the top-k have actually arised. The pseudocode is given up Formula 1.

Algorithm 1 CMiner

Input: Set of items I, Item of interest i ∈ I, feature space F, Collection Q ∈ 2F of queries with non-zero weights, skyline pyramid DI, int k

Output: Set of top-k competitors for i

1: TopK ← masters(i)
2: **if** (k ≤ |TopK|) **then**
3: **return** TopK
4: **end if**
5: k ← k − |TopK|
6: LB ← −1
7: X ←GETSLAVES(TopK;DI) ∪ DI[0]
8: **while** (|X | != 0) **do**
9: X UPDATETOPK(k;LB;X)
10: **if** (|X | != 0) **then**
11: TopK MERGE(TopK;X)
12: **if** (|TopK| = k) **then**
13: LB WORSTIN(TopK)
14: **end if**
15: X GETSLAVES(X;DI)
16: **end if**
17: **end while**
18: **return** TopK
19: **Routine** UPDATETOPK(k, LB, X)
20: localTopK ← ∅
21: low(j) ← 0; ∀j ∈ X.
22: up(j) ←
Σ
q2Q
p(q) × V q
j;j ; ∀j ∈ X.
23: **for** every q ∈ Q **do**
24: maxV ← p(q) × V q
i;i
25: **for** every item j ∈ X **do**
26: up(j) ← up(j) − maxV + p(q) × V q
i;j
27: **if** (up(j) < LB) **then**
28: X ← X \ {j}
29: **else**
30: low(j) ← low(j) + p(q) × V q
i;j
31: localTopK:update(j; low(j))
32: **if** (|localTopK| ≥ k) **then**
33: LB ←WORSTIN(localTopK)
34: **end if**
35: **end if**
36: **end for**
37: **if** (|X | ≤ k) **then**
38: **break**
39: **end if**

40: **end for**

41: **for** every item j ∈ X **do**

42: **for** every remaining q ∈ Q **do**

43: low(j) ← low(j) + p(q) × V q

i;j

44: **end for**

45: localTopK:update(j; low(j))

46: **end for**

47: **return** TOPK(localTopK)

IV. EXPERIMENTAL EVALUATION

Datasets and Baselines: Our experiments consist of 4 datasets, which were gathered for the objectives of this job. The datasets were purposefully picked from various domain names to depict the cross-domain applicability of our strategy. Along with the complete details on each thing in our datasets, we additionally accumulated the complete collection of evaluations that were readily available on the resource internet site. These testimonials were made use of to (1) quote inquires chances, as defined in Area 2.2 as well as (2) draw out the point of views of customers on certain features.The highly-cited technique by Ding et al. [28] is utilized to transform each testimonial to a vector of point of views, where each point of view is specified as a feature-polarity mix (e.g. solution+, food-). The portion of evaluations on a thing that share a favorable viewpoint on a particular attribute is utilized as the function's numerical worth for that product. We describe these as point of view attributes. Table 4 consists of detailed data for each and every dataset, while a comprehensive summary is supplied listed below. ELECTRONIC CAMERAS: This dataset consists of 579 electronic cams from Amazon.com. We gathered the complete collection of testimonials for every cam, for an overall of 147192 evaluations. The collection of attributes consists of the resolution (in MP), shutter rate (in secs), zoom (e.g. 4x), as well as rate. It likewise consists of viewpoint attributes on handbook, pictures, video clip, layout, flash, emphasis, food selection alternatives, lcd display, dimension, attributes, lens, guarantee, shades, stablizing, battery life, resolution, as well as expense. RESORTS: This dataset consists of 80799 testimonials on 1283 resorts from Booking.com. The collection of attributes consists of the centers, tasks, and also solutions provided by the resort. All 3 of these multi-categorical functions are readily available on the internet site. The dataset additionally consists of point of view functions on area, solutions, tidiness, personnel, as well as convenience.

Dataset	#Items	#Feats.	#Subsets	Skyline Layers
CAMERAS	579	21	14779	5
HOTELS	1283	8	127	5
RESTAURANTS	4622	8	64	12
RECIPES	100000	22	133	22

Table 1: Dataset Statistics

For each and every dataset, the second, third, fourth and also fifth columns consist of the variety of products, the variety of attributes, the variety of unique questions, and also the variety of layers in the particular sky line pyramid, specifically. In order to end the summary of our datasets, we provide some data on the skyline-pyramid framework built for each and every corpus. Number 4 reveals the circulation of things in the very first 6 sky line layers of each dataset. We observe that, for all datasets, virtually 99% of the things can be discovered within the initial 4 layers, with most of those dropping within the very first 2 layers. This is because of the big dimensionality of the attribute area, that makes it hard for things to control each other. As we display in our experiments, the horizon pyramid allows CMiner to plainly surpass the standards relative to computational price. This is regardless of the high focus of things within the very first layers, considering that CMiner can properly go across the pyramid as well as take into consideration just a little portion of these things.

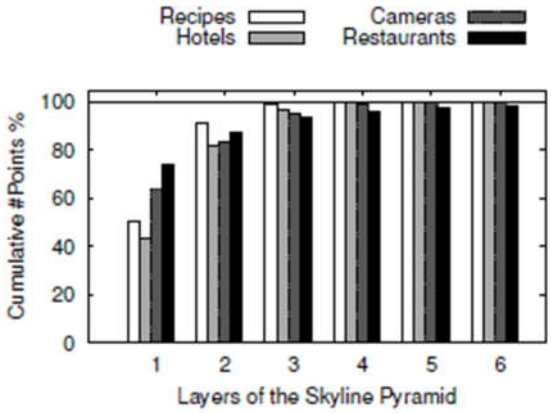

Fig 4: Cumulative distribution of items across the first 6 layers of the skyline pyramid.

Pruning Efficiency: Much of CMiner's performance comes from its capability to dispose of or precisely examine prospects. We show this in Number 7. The number consists of one collection of bars for every dataset, with each bar standing for a various worth of k (k 2 f3, 10, 50, 150, 300g, in the order revealed). The white section of each bar (post-pruned) stands for the ordinary variety of products trimmed within UPDATETOPK() (line 28). There, a product is trimmed if, as we discuss the collection of inquiries Q, its top bound gets to a worth that is less than POUND (the most affordable rival in the present top-K). The black section of each bar (pre-pruned) stands

for the typical variety of products that were never ever included in the prospect collection X due to the fact that their best-case circumstance (self insurance coverage) was apriori even worse than POUND. As a result, they can be removed and also we do not need to consider their competition in the context of the questions. We clarify this system carefully in the last paragraph of Area 4.2. Ultimately, the patternfilled part (unpruned) on top of each bar describes the ordinary variety of things that were completely assessed in their whole (i.e. for all questions). We observe that the large bulk of prospects is removed by among both kinds of trimming that we think about below. The high variety of prepruned inquiries is especially motivating, as it indicates the highest possible computational financial savings. Lastly, it is very important to keep in mind that these searchings for correspond throughout datasets.

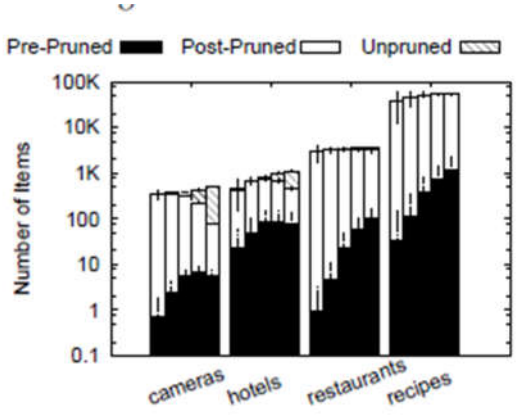

Fig 5: Pruning Effectiveness

V. CONCLUSION

We offered an official meaning of competition in between 2 products, which we verified both quantitatively as well as qualitatively. Our formalization applies throughout domain names, conquering the imperfections of previous techniques. We think about a variety of aspects that have actually been greatly neglected in the past, such as the placement of the things in the multi-dimensional attribute area as well as the choices and also point of views of the individuals. Our job presents an end-to-end technique for extracting such info from big datasets of client testimonials. Based upon our competition interpretation, we attended to the computationally tough trouble of locating the top-k rivals of a provided thing. The recommended structure is reliable as well as appropriate to domain names with huge populaces of things. The effectiveness

of our approach was confirmed through a speculative examination on actual datasets from various domain names. Our experiments likewise exposed that just a handful of testimonials suffices to with confidence approximate the various kinds of individuals in an offered market, too the variety of customers that come from each kind.

VI. REFERENCES

[1] W. T. Few, "Managerial competitor identification: Integrating the categorization, economic and organizational identity perspectives," Doctoral Dissertaion, 2007.

[2] M. Bergen and M. A. Peteraf, "Competitor identification and competitor analysis: a broad-based managerial approach," Managerial and Decision Economics, 2002.

[3] J. F. Porac and H. Thomas, "Taxonomic mental models in competitor definition," The Academy of Management Review, 2008.

[4] M.-J. Chen, "Competitor analysis and interfirm rivalry: Toward a theoretical integration," Academy of Management Review, 1996.

[5] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in ICDM, 2006.

[6] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relationships from online news: A network-based approach," Electronic Commerce Research and Applications, 2011.

[7] R. Li, S. Bao, J. Wang, Y. Liu, and Y. Yu, "Web scale competitor discovery using mutual information," in ADMA, 2006.

[8] S. Bao, R. Li, Y. Yu, and Y. Cao, "Competitor mining with the web," IEEE Trans. Knowl. Data Eng., 2008.

[9] G. Pant and O. R. L. Sheng, "Avoiding the blind spots: Competitor identification using web text and linkage structure," in ICIS, 2009.

[10] D. Zelenko and O. Semin, "Automatic competitor identification from public information sources," International Journal of Computational Intelligence and Applications, 2002.

[11] R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing, vol. 27, no. 4, pp. 293–307, 2010.

[12] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, 1980.

[13] R. Deshpand and H. Gatingon, "Competitive analysis," Marketing Letters, 1994.

[14] B. H. Clark and D. B. Montgomery, "Managerial Identification of Competitors," Journal of Marketing, 1999.

SHANMUKHI SAI NAINAR: BTech Student, RMK college of engineering, RSM nagar, Gummadipoondi Taluk, Tiruvallur district, Kavaraipettai, Tamilnadu -601206, India.

YARRAGUNTLA SAISWETHA: BTech Student, R V R & J C College of Engineering, Chandramoulipuram, Chowdavaram, Guntur-522 019, Andhra Pradesh, India