International Journal of Research in Engineering, Science and Management
Volume-2, Issue-8, August-2019
www.ijresm.com | ISSN (Online): 2581-5792

Mining Competitors from Large Unstructured Datasets

D. Balachandar[1], S. Amaresan[2]
[1]*Research Scholar, Department of Computer Science, Ponnaiyah Ramajayam Institute of Science & Technology (PRIST University), Thanjavur, India*
[2]*Associate Professor, PRIST University, Thanjavur, India*

***Abstract*: Data mining is that the method of sorting through giant information sets to spot patterns and establish relationship to resolve through data analysis.**

One of the major unsolved problems is the management of unstructured data. The unstructured data such as multimedia files, documents, comments, customer support request, news, emails, reports and web pages are difficult to capture and store in the common database system. Data mining is the popular area of the research which facilitates the business improvement process such as mining user preference, mining web information to get opinion about the product or services. In the current competitive business scenario, there is a need to analyse the competitive features and factors of an item that most affect its competitiveness. The evaluation of competitiveness always uses the customer opinions in terms of reviews, ratings and abundant source of information from the web and other sources. The challenges arise on the features of main competitors of a given item. In this paper we proposed cminer with condition Artificial intelligence (CAI) to for evaluating competitiveness in large datasets and addresses the natural problem finding top-level competitors.

***Keywords*: Data mining, unstructured data, competitive features, Artificial intelligence, Cminer.**

1. Introduction

The strategic importance of detecting and observing business competitors is an inevitable research, which motivated by several business challenges. Monitoring and identifying firm's competitors have studied in the earlier work. Data mining is the optimal way of handling such huge information's for mining competitors. Item reviews form online offer rich information about customers' opinions and interest to get a general idea regarding competitors. However, it is generally difficult to understand all reviews in different websites for competitive products and obtain insightful suggestions manually. In the earlier works in the literatures, many authors analyzed such big customer data intelligently and efficiently [1]-[3]. For example, a lot of studies about online reviews were stated to gather item opinion analysis from online reviews in different levels. However, most researchers in this field ignore how to make their findings be seamlessly utilized to the competitor mining process.

Competitive intelligence initially classifies the potential risk and chances by collecting the information about the context to handle the manager in making tactical decisions for an organization. Many organization recognizes the significance of competitive intelligence in enterprise risk management and decision support system. They also invest a great amount of money in competitive intelligence. The fundamental significance of customer choices, e.g., in correlation with new product expansion procedures. These procedures are broadly affirmed in marketing research. Usually customer choices are evaluated through conjoint analysis using online or paper-pencil survey. Though, this type of choices can highly price with reference to time and money [4].

Existing research based on mining comparative articulations (e.g. "product A is superior than product B") from the online or alternative documentary sources [5]-[7].

However, this articulation can certainly be sign of competitiveness and they are missing in numerous domains. For example, while competing brand names at the company level (e.g. Google vs Yahoo or Sony vs. Panasonic). While comparing these patterns, it can be found by simply questioning on the web. But, it is easy to classify mainstream domains where such facts are tremendously uncommon, such as jewelry, hotels, restaurants and furniture. Inspired by these limitations, we present a new description of the competitiveness between two items on the basis of market sectors.

2. Artificial Neural Network

An artificial neural network (ANN), usually called neural network.

A. Related work

R. Decker et. al. [8] discussed on customer reviews that are widely accessible on the internet for huge number of product classifications. The experts and cons articulated in this manner discover individually observed the strengths and weaknesses of the particular products, while the typically allocated product rankings represent their complete valuation. The important query from this point is by what means to turn the accessible plentitude of individual customer opinions into aggregate customer first choice which can be utilize in product expansion or improvement procedures. To overcome these shortcomings author presents an econometric framework that can be utilized

to the cited type of data and used natural language processing methodologies. The suggested procedure simplifies the evaluation of parameters allow implications on the comparative outcome of product features and brand names on the complete evaluation of the products.

C. W.-K. Leung et. al. [9] proposes a novel Probabilistic Rating Inference Framework, well known as PREF, for mining user choices from reviews and then map out such choices onto numerical rating scales. PREF utilizes existing linguistic processing methods to extract opinion words and product attributes from reviews. It then estimates the sentimental orientations (SO) and strength of the opinion words by means of our proposed relative-frequency-based technique. In this paper, author presents a Preference technique i.e. a probabilistic rating inference framework. Pref is a probabilistic rating inference framework model which helps to developed and to support the integration of sentiment analysis and CF.

It includes four steps: 1) Data preparation: This method processes the user opinions for the subsequent analysis. Different preprocessing methods may be needed to depend on the data sources.

E. Marrese-Taylor et. al. [10] present and extend an approach of Bing Liu's aspect-based opinion mining methodology to utilize it to the tourism domain. Author also suggested a method for considering a novel alternative to uncover customer opinions regarding tourism products, specifically hotels and restaurants using opinions accessible on the web as reviews. To estimate this suggestion, author also conducted an experiment using hotel and restaurant reviews found from Trip Advisor. This outcome displayed that tourism product reviews available on web sites comprise valuable information about customer opinions that can be extracted using an aspect-based opinion mining method.

K. Lerman et. al. [11] presents endeavors to achieve a novel assignment of mining focused data regarding an element, the element, for example, an organization, item or individual from the web. The creators proposed a calculation called "CoMiner", which initially separates an arrangement of similar applicants of the information substance and afterward positions them as indicated by the likeness, lastly extricates the focused fields. In any case, the CoMiner particularly created to help for particular space. However, the exertion for the further spaces is as yet difficult.

3. Methodology

The proposed 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, including an efficient evaluation algorithm and an appropriate index. fusion of Artificial Neural Networks (ANN) and CMiner have attracted the growing interest of researchers in numerous scientific and engineering areas because of the growing would like of adaptive intelligent systems to unravel the problems associated with e-commerce.

ANN learns from scratch by adjusting the interconnections between layers. NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. It consists of an interconnection of simple components referred to as neurons, which are programming constructs that mimic the properties of biological neurons. ANNs comprises one or additional layers. Each layer has one or additional neurons. The neuron (perceptron) can be defined simply as a device with many inputs, one output, and an activation function. The neurons are connected by directed links. Each link has a numeric weight associated with it from which weighted sum can be calculated. Then after classes can be distinguished by using activation function.

4. Cminer algorithm

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 $\mathcal{I}$, Item of interest $i \in \mathcal{I}$, feature space $\mathcal{F}$, Collection $\mathcal{Q} \in 2^{\mathcal{F}}$ of queries with non-zero weights, skyline pyramid $\mathcal{D}_{\mathcal{I}}$, int k
Output: Set of top-k competitors for i

$TopK \leftarrow masters(i)$
if ($k \leq |TopK|$) **then**
  return $TopK$
end if
$k \leftarrow k - |TopK|$
$LB \leftarrow -1$
$\mathcal{X} \leftarrow$ GETSLAVES($TopK, \mathcal{D}_{\mathcal{I}}$) $\cup\, \mathcal{D}_{\mathcal{I}}[0]$
while ($|\mathcal{X}|$!= 0) **do**
  $\mathcal{X} \leftarrow$ UPDATETOPK($k, LB, \mathcal{X}$)
  if ($|\mathcal{X}|$!= 0) **then**
    $TopK \leftarrow$ MERGE($TopK, \mathcal{X}$)
    if ($|TopK| = k$) **then**
      $LB \leftarrow$ WORSTIN($TopK$)
    end if
    $\mathcal{X} \leftarrow$ GETSLAVES($\mathcal{X}, \mathcal{D}_{\mathcal{I}}$)
  end if
end while
return $TopK$

Routine UPDATETOPK(k, LB, $\mathcal{X}$)
$localTopK \leftarrow \emptyset$
$low(j) \leftarrow 0, \forall j \in \mathcal{X}.$
$up(j) \leftarrow \sum_{q \in \mathcal{Q}} p(q) \times V^q_{j,j}, \forall j \in \mathcal{X}.$
for every $q \in Q$ **do**
  $maxV \leftarrow p(q) \times V^q_{i,i}$
  for every item $j \in \mathcal{X}$ **do**
    $up(j) \leftarrow up(j) - maxV + p(q) \times V^q_{i,j}$
    if ($up(j) < LB$) **then**
      $\mathcal{X} \leftarrow \mathcal{X} \setminus \{j\}$
    else
      $low(j) \leftarrow low(j) + p(q) \times V^q_{i,j}$
      $localTopK.update(j, low(j))$
      if ($|localTopK| \geq k$) **then**
        $LB \leftarrow$ WORSTIN($localTopK$)
      end if
    end if
  end for
  if ($|\mathcal{X}| \leq k$) **then**
    break
  end if
end for
for every item $j \in \mathcal{X}$ **do**
  for every remaining $q \in \mathcal{Q}$ **do**
    $low(j) \leftarrow low(j) + p(q) \times V^q_{i,j}$
  end for
  $localTopK.update(j, low(j))$
end for
return TOPK($localTopK$)

5. Results and discussion

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.

Table 1 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, stabilizing, 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.

Table 1
Dataset statistics

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

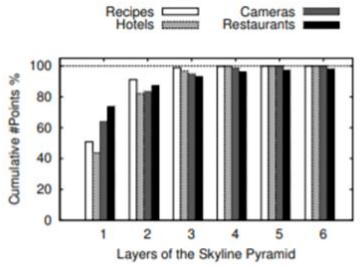

Fig. 1. Layers of skyline pyramid

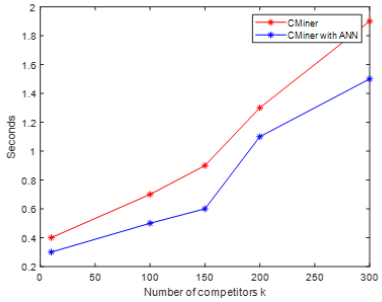

Fig. 2. Average time (per item) to compute top-k competitors for Cameras dataset

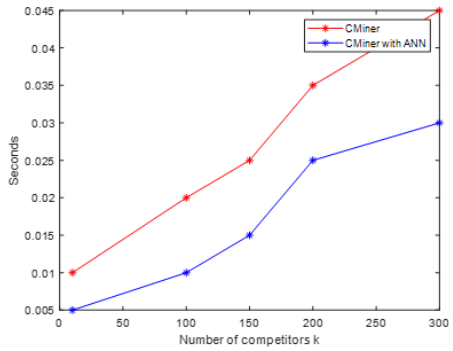

Fig. 3. Average time (per item) to compute top-k competitors for Hotels dataset

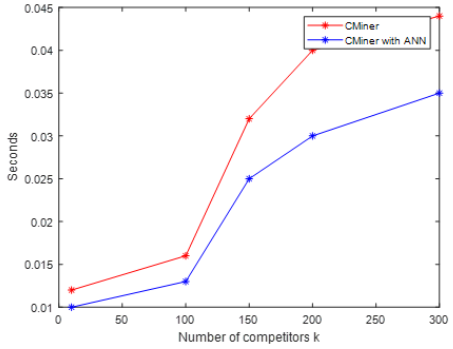

Fig. 4. Average time (per item) to compute top-k competitors for Restaurants dataset

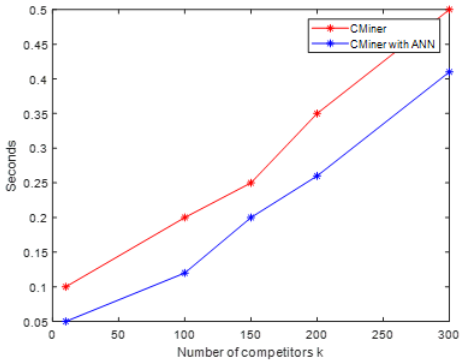

Fig. 5. Average time (per item) to compute top-k competitors for Receipes dataset

6. Conclusion

Data mining has significance with respect to finding the examples, estimating, disclosure of learning and so forth, in various business areas. Machine learning methodologies are broadly utilized as a part of different applications. Each business-related application utilizes information mining systems. To enhance such business or giving proper competitor to the business to the client require the help of web mining systems. The competitor mining is one such an approach to investigate competitors for the preferred items. 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.

References

[1] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, 1980.
[2] R. Deshpand and H. Gatingon, "Competitive analysis," Marketing Letters, 1994.
[3] B. H. Clark and D. B. Montgomery, "Managerial Identification of Competitors," Journal of Marketing, 1999.
[4] K. Xu, S. S. Liao, J. Li, and Y. Song, "Mining comparative opinions from customer reviews for competitive intelligence," Decis. Support Syst., 2011.
[5] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining rival relationships from on-line news: A network-based approach," Electronic Commerce Research and Applications, 2011.
[6] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in ICDM, 2006.
[7] 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.
[8] 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.
[9] C. W. K. Leung, S. C. F. Chan, F. L. Chung, and G. Ngai, "A probabilistic rating reasoning framework for mining user preferences from reviews," World Wide Web, vol. 14, no. 2, pp. 187–215, 2011.
[10] E. Marrese-Taylor, J. D. Vel´asquez, F. Bravo-Marquez, and Y. Matsuo, "Identifying client preferences concerning commercial enterprise merchandise mistreatment AN aspect-based opinion mining approach," Procedia engineering, vol. 22, pp. 182–191, 2013.
[11] K. Lerman, S. Blair-Goldensohn, and R. McDonald, "Sentiment summarization: evaluating and learning user preferences," in ACL, 2009, pp. 514–522.