16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

A Bipartite Graph Based Competitiveness Degrees Analysis with Query Logs on Search Engine

Dandan Qiao[1] Qiang Wei[1] Jin Zhang[2] Guoqing Chen[1]

[1] School of Economics and Management, Tsinghua University, Beijing 100084, China
{qiaodd.12, weiq, chengq}@sem.tsinghua.edu.cn
[2] School of Business, Renmin University of China, Beijing 100872, China, zhangjin@rbs.org.cn

Abstract

Competitive intelligence analysis based on user generated contents (UGCs) shows advantages on possessing the benefit of the wisdom of crowds, evading cognitive biases and timely updating. This paper investigates and constructs a bipartite graph model by extracting joint co-occurrence of competitors from query logs, i.e., an important type of UGCs. The model represents the inherent competitiveness among competitors, based on which their mutual competitiveness degrees can be measured. Moreover, the *BGComp* algorithm is designed as well. Finally, an analysis on forecasting automobiles' sales demonstrates the better forecasting power of the regression models integrating the competitiveness degrees than the benchmark model.

Keywords: Competitiveness degree, Query log, Bipartite graph model

1. Introduction

It is widely recognized that competitive intelligence analysis serves as an important role in companies for strategy formulation, implementation, monitoring and adjustment [1]. Generally, there exist two major stages in competitive intelligence analysis, i.e., competitor identification and competitiveness degree analysis [2, 3]. With competitor identification, companies who have a full understanding of the competitors can be well positioned to achieve competitive advantage [3]. Furthermore, Competitiveness degree analysis helps managers quantify the degrees of competition with their competitors [4, 5]. Besides serving as a pivotal role in the strategy management, competition would also significantly influence a company's market performance, e.g., sales, market shares, etc. [6], since each move the company takes (e.g., offering new products or promoting new services) would elicit various degrees of responses from its competitors, vice versa [7]. Therefore, when analyzing and predicting the market performance of a company/product, after identifying its competitors, to what degrees that it competes with the competitors must be taken into consideration.

In recent years, the studies of competitive intelligence analysis have been broadly fuelled by the flourish marketing activities. Specifically, the studies on competitive intelligence analysis is conducted from two streams, i.e., attribute-rating based [4, 8] and consideration-set based [9, 10]. In attribute-rating based analysis, it's commonly believed that the utility of a company/product/brand is perceived as an agglomeration of selected attributes by managers/experts. Then the competitiveness can be further analyzed based on the set of agglomerated attributes. Clearly, this stream highly depends on managers/experts' judgments on attributes enumeration and evaluation, which increases the difficulty, especially in a hyper-competitive environment where consumers' appetites change rapidly and new competitors emerge frequently [11]. Besides, attributes from managers are probably inconsistent with consumers' perceived attributes, which are hard to define and observe directly. The later stream contends that companies/products/brands captured in the same consideration set by consumers are de facto competitors [10]. But there usually exists a ceiling for consumers' memory, i.e., limited rationality, leading to a restricted size of consideration set to six or seven instances [12, 13]. Besides, both two streams exhibit obvious cognitive biases existed in the limited quantity of self-report surveys [14, 15], e.g., usually only dozens of experts or hundreds of consumers can be surveyed, which is a quite small sample against the whole market size. Moreover, due to the non-structuration and low-automation, the analytical process is usually tedious and time-consuming, leading to deferred response to market dynamics.

Nowadays, facing the colossal quantity of Internet data, it has been recently observed that the frequently updated online user generated contents (UGCs) can be regarded as rich and timely resources for analyzing competitive intelligence [15, 16]. For example, consumers' online reviews reflect their preference on different rivals; users' search logs with certain products/brands usually contain their highlighted attributes; online news contains the comments and descriptions on a group of substitutes, etc. Therefore, due to the Big Data characteristics of UGCs, analyzing with UGCs can greatly overcome the shortcomings of the traditional methods on threefold. First, UGCs are generated by users actively, which can to a large extent evade the cognitive biases in self-report format. Second, the UGCs created by the large volume of users possess the benefit of "the wisdom of crowds" and are more likely to break the ceiling of limited consideration set after aggregation. Third, timely and dynamically updated UGCs potentially imply up-to-date competitive intelligence.

Focusing on different types of online UGCs, some efforts have been conducted to analyze competitive intelligence, which will be briefly reviewed in the Related Work section. However, the effectiveness of these works highly depends on the UGCs containing co-occurred competitors' information, i.e., two or multiple

competitors' information have to appear in the same texts, web pages, snippets, etc., which could be scarce sometimes [16] and significantly restricts their applicability. Besides, due to the format variety of UGCs, e.g., webpages, snippets, news, links, reviews, etc., the analyzing techniques are also complicated to some extent. Furthermore, little work is conducted on measuring the competitiveness degree among competitors by leveraging UGCs, which, however, is quite important for market performance evaluation and forecasting.

For measuring competitiveness degrees, a group of competitors of concern are usually identified in advance, called focused competitors. Thus, this paper will propose a method to measure the competitiveness degrees among the focused competitors, by constructing a bipartite graph model with extracted joint co-occurrence information from query logs on search engine. The proposed method shows advantages on the three aspects. First, query logs as a typical rich source of UGCs, contain search engine users' intents and have relatively more structural format compared with other UGCs, e.g., web pages, links, reviews, etc., which is more suitable for large-scale analysis. Second, the bipartite graph model constructed by extracting joint co-occurrence information in query logs can reflect inherent perceptions on competitors of massive search engine users. Third, the competitiveness degree among competitors can be measured and further utilized to help companies understand the market's competitive dynamics in depth. To the most knowledge, this study is the first attempt to measure competitiveness degrees based on query logs.

This paper is organized as follows. Section 2 briefly overviews the related work. Section 3 introduces how to construct the bipartite graph model with a query log and defines the notion of competitiveness degree. Section 4 proposes an algorithm for constructing a bipartite graph model by extracting information from a query log and further calculating competitiveness degrees. A real data experiment using derived competitiveness degrees to improve sales forecasting is illustrated in Section 5, showing the outperformance of the proposed method. Section 6 presents the final conclusion.

2. Related Work

The availability of rich UGCs online has been fuelling the emphasis on competitive intelligence analysis by using different digital resources [17]. Bao et al. proposed an algorithm CoMiner based on the co-occurrence of competing brands in search engine results [18]. Ma et al. built a network-based method to predict competitive relationships using the co-occurrence in news stories [19]. Xu proposed a graphical model to extract competitive relationships from customer views [20]. And Pant et al. presented a method to find competitors by analyzing text and linkage structure of the relevant pages on the web [15]. These studies have greatly promoted the competitiveness-driven applications using online UGCs. However, they are mostly conditioned on the premise that competitors have much more co-occurrence in the information sources. Such co-occurrence oriented comparative evidence doesn't always work well due to the fact that scarce co-occurrence information could be directly retrieved in UGCs. Besides, competitor identification is just the precursor for further competition analysis [14]. There lacks an appropriate definition of competitiveness degree to help measure the extent to which two competitors compete with each other in UGCs. Though recently some scholars attempted to quantify the competitiveness from products/companies /brands' attributes [8, 16], of which the drawbacks have been specified in the discussion of attribute-rating based methods in the Introduction section, i.e., difficulty in enumeration and evaluation on attributes due to managers/experts' cognitive biases. Thus, the research of competitiveness degree measuring with UGCs is still quite limited.

Query logs are believed to reflect a large majority of users' intents [21, 22], and are more structural to process, i.e., in form of conjunctive keywords with corresponding query volumes, bringing abundant researches to explore the relationships between search data and diverse social phenomena, e.g., the unemployment rate prediction [23], stock prices prediction [24], etc. Furthermore, Choi and Varian also demonstrated the significant value of query logs for nowcasting the automotive sales [25]. These various applications of query logs claim that the search volume of query keywords about a certain variable in question could be used to measure its corresponding market shares/sales. Nevertheless, little work has been conducted on competitiveness degree analysis with query log data, since seldom do two competitors co-occur in the same query.

This paper will further investigate the value deeply hidden in query logs. Though two competitors can hardly co-occur in the same queries, two competitors are highly possible to co-occur with the same intermediate keywords in different queries, called joint co-occurrence. For example, "*BMW*" hardly co-occurs with "*Audi*", but "*BMW Germany*" and "*Audi Germany*", i.e., "*Germany*" as an intermediate, could be frequently observed. Further, more intermediate keywords, e.g., "*car*", "*repair*", "*insurance*", etc., could be observed to frequently co-occur with either "*BMW*" or "*Audi*". Therefore, these intermediate keywords could be regarded as a certain of "perceived attributes" of the focused competitors, reflecting users' perception, called attributes in this paper. With the Big Data characteristic of query logs, based on the crowded wisdom of search engine users, the competitiveness degrees could be measured by analyzing the joint co-occurrence among the focused competitors in query logs.

3. A Bipartite Graph Based Competitiveness Model

As discussed in previous section, a co-occurred keyword for a certain focused competitor could be deemed as an attribute of the competitor, with the corresponding query volume indicating the cognitive strength of all users on a search engine. Moreover, the attribute may also co-occur with other focused competitors with different query volumes.

Given a set of focused competitors C, a corresponding query log Q is a set containing n queries, i.e., for

each $q \in Q$, q is denoted as a triple (c, a, vol), where $c \in C$, a is an attribute, and vol is the volume of querying c and a simultaneously. A real query log example on focused competitors' keywords "*Infiniti*", "*BMW*" and "*Audi*" from Google is like {(*Infiniti*, *cars*, 23,100), (*Infiniti*, *car insurance*, 390), (*Infiniti*, *SUV*, 27,100), (*BMW*, *cars*, 22,200), (*BMW*, *Germany*, 1,600), (*BMW*, *SUV*, 22,200), (*BMW*, *X5*, 9,880), (*Audi*, *cars*, 18,100), (*Audi*, *Germany*, 1,300), (*Audi*, *SUV*, 27,100), (*Audi*, *Q5*, 6,773), …}. After allocating all focused competitors as well as attributes, a bipartite graph model could be constructed as follows.

A bipartite graph model is denoted as G, where $G = (C, A, Q)$. C is the set of n focused competitors, i.e., $C = \{c_1, c_2, \ldots, c_n\}$. A is the set of m attributes, i.e., $A = \{a_1, a_2, \ldots, a_m\}$, which can be derived based on their co-occurrence with all focused competitors. Q is the set of corresponding queries q_{ij}, i.e., $q_{ij} = (c_i, a_j, vol_{ij})$, where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Figure 1 depicts a bipartite graph model example for the focused competitors "*BMW*", "*Infiniti*" and "*Audi*" with crawled date from Google. In Figure 1, the nodes named "*Infiniti*", "*BMW*" and "*Audi*" represent the focused competitors and the nodes labelled as "*Cars*", "*Germany*", etc., are attributes with respect to the focused competitors. In addition, a particular edge in Figure 1 represents the query with the focused competitor and an attribute. The number located near the edge is the corresponding volume of the query.

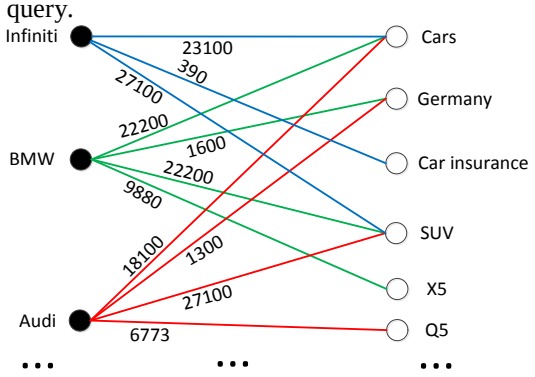

Fig. 1: An example of bipartite graph model

Figure 1 clearly indicates that, though the focused competitors hardly co-occur in the same queries, their competitiveness could be measured through the paths via different intermediate attributes, i.e., forming joint co-occurrences.

For a focused competitor c_x, $x = 1, 2, ..., n$, without loss of generality, c_x co-occurs with a_j, $j = 1, 2, \ldots, m$, with vol_{xj} ($vol_{xj} = 0$, if c_x and a_j do not co-occur) in the query log. Thus, the probability that a_j co-occurs with c_x, denoted as $p(c_x \rightarrow a_j)$, can be measured as follows:

$$p(c_x \rightarrow a_j) = \frac{vol_{xj}}{\sum_{s=1}^{m} vol_{xs}}. \quad (1)$$

Semantically, $p(c_x \rightarrow a_j)$ reflects the degree that search engine users deem a_j as an attribute of c_x, since $p(c_x \rightarrow a_j)$ of the users conjunctively query a_j with c_x.

Further, for an attribute a_j, $j = 1, 2, \ldots, m$, the probability that c_y co-occurs with a_j, denoted as $p(a_j \rightarrow c_y)$, can be measured as follows:

$$p(a_j \rightarrow c_y) = \frac{vol_{yj}}{\sum_{t=1}^{n} vol_{tj}}. \quad (2)$$

Similarly, $p(a_j \rightarrow c_y)$ reflects the degree that search engine users deem c_y as a candidate competitor in their consideration set while concerning attribute a_j, since $p(a_j \rightarrow c_y)$ of the users conjunctively query c_y with a_j.

Thus, it can be inferred that the joint degree $p(c_x \rightarrow a_j) \times p(a_j \rightarrow c_y)$ reflects to what extent that c_y could be recognized as a competitor of c_x by search engine users concerning attribute a_j, i.e., the degree of c_y competing with c_x via a_j, denoted as $Comp_{aj}(c_x \rightarrow c_y)$.

For illustrative purpose, suppose a bipartite graph model containing two focused competitors, i.e., c_x and c_y, and only one single co-occurring attribute a, with vol_{xa} and vol_{ya}, respectively. Thus, $Comp_a(c_x \rightarrow c_y) = vol_{xa}/vol_{xa} \times vol_{ya}/(vol_{xa} + vol_{ya}) = vol_{ya}/(vol_{xa} + vol_{ya})$, and $Comp_a(c_y \rightarrow c_x) = vol_{xa}/(vol_{xa} + vol_{ya}) \neq Comp_a(c_x \rightarrow c_y)$. Moreover, $Comp_a(c_x \rightarrow c_y) + Comp_a(c_y \rightarrow c_x) = 100\%$ and $Comp_a(c_x \rightarrow c_x) = vol_{xa}/(vol_{xa} + vol_{ya})$. Intuitively, taking a search engine as a market of keywords with corresponding query volumes as their sales, $Comp_a(c_x \rightarrow c_y)$ is consistent with the market share of c_y, demonstrating the competitiveness of c_y on c_x, vice versa. Besides, $Comp_a(c_x \rightarrow c_x)$ is just the market share of c_x itself, which is also consistent with common knowledge.

Since usually multiple attributes co-occur with c_x and c_y in the bipartite graph model, after aggregating with all attributes, the aggregated competitiveness degree that c_y competes with c_x can be defined.

Definition 1: Given a $G = (C, A, Q)$, for any two competitors c_x and c_y, where c_x, $c_y \in C$, pairwise competitiveness degree that c_y competes with c_x, denoted as $Comp(c_x \rightarrow c_y)$ or $Comp_{xy}$ in short, can be defined as:

$$Comp(c_x \rightarrow c_y) = \sum_{j=1}^{m} p(c_x \rightarrow a_j) \times p(a_j \rightarrow c_y). \quad (3)$$

Clearly, $Comp(c_y \rightarrow c_x)$ represents the agglomeration of degrees on multiple attributes via which c_y competes with c_x. It can be found that $Comp(c_y \rightarrow c_x) \neq Comp(c_x \rightarrow c_y)$. Moreover, an important property could be further derived.

Property 1: Given a $G = (C, A, Q)$ and derived competiveness degrees, for a certain c_x, $c_x \in C$, the following property holds:

$$\sum_{y=1}^{n} Comp(c_x \rightarrow c_y) = 100\%. \quad (4)$$

The proof is omitted here to save the space. Property 1 indicates that the summarized competitiveness degrees of all competitors with respect to a given competitor is 100%, demonstrating the consistency that the defined competitiveness degrees could be regarded as market shares of competitors' keywords perceived by users on a search engine.

Thus, given a set of focused competitors, the mutual competitiveness degrees among competitors can be calculated with the constructed bipartite graph model on query logs. The corresponding calculating algorithm, called *BGComp*, will be proposed in next section.

4. The BGComp Algorithm

With a given set of focused competitors C, in order to calculating the mutual competitiveness degrees among

all the focused competitors, the queries on a search engine, e.g., Google, should be preprocessed and screened to retrieve the query log containing focused competitors, i.e., deriving Q. This surely is a computation-intensive operation. Then, based on C and the derived Q, the queries in Q need to be scanned and the corresponding co-occurred attributes can be retrieved to form the set A. In this step, some basic cleaning operations should be performed to remove noises and disturbance. Thereafter, the co-occurrence between each competitor and a corresponding attribute can be denoted as an edge with its corresponding query volume, thus the bipartite graph model G can be constructed.

Based on the constructed G, in order to calculate the mutual competitiveness degrees of all focused competitors, the following algorithmic strategy is adopted. First, for each competitor c_x in C and all attributes in A, their $p(c_x \rightarrow a_j)$s, $j = 1, 2, \ldots, m$, are calculated. Next, for each attribute a_j in A and all competitors in C, their $p(a_j \rightarrow c_y)$s, $y = 1, 2, \ldots, n$, are calculated. Finally, based on the derived degrees, for any two competitors, e.g., c_x and c_y, the competitiveness degree can be calculated with formula (3). Figure 2 lists the algorithmic sketch.

BGComp Algorithm: A Bipartite Graph Model Based Competitiveness Degree Calculation

Input: $C = \{c_1, c_2, \ldots, c_n\}$;
Q_0 = Original query log from search engine
Output: $Comp(c_x \rightarrow c_y)$, $\forall c_x, c_y \in C$
Initialization: $A = \varnothing$; $Q = \varnothing$;

```
1.   Q = Preprocess(Q0, C);                    // Deriving Q
2.   A = Extract(Q, C);                        // Constructing A
3.   for ∀cx ∈ C do                           // Calculating p(cx→aj)
4.   {          for ∀aj ∈ A do
5.              {          Calculate p(cx→aj);    }
6.   }
7.   for ∀aj ∈ A do                           // Calculating p(aj→cy)
8.   {          for ∀cy ∈ C do
9.              {          Calculate p(aj→cy);    }
10.  }
11.  for ∀cx ∈ C do                           // Calculating Comp(cx→cy)
12.  {          for ∀cy ∈ C do
13.             {          Calculate Comp(cx→cy);          }
14.  }
```

Fig. 2: Algorithmic sketch of BGComp

From Figure 2, it could be found that the *BGComp* algorithm is composed of five major steps, i.e., (1) preprocessing query log Q (line 1), (2) extracting the set of attributes A (line 2); (3) calculating $p(c_x \rightarrow a_j)$s (lines 3-6), (4) calculating $p(a_j \rightarrow c_y)$s (lines 7-10) and (5) calculating $Comp(c_x \rightarrow c_y)$s (lines 11-14). Assume that the size of Q is l, the number of focused competitors is n, and the number of extracted attributes is m. It should be emphasized that m cannot be configured exogenously but extracted from Q according to C. Thus step (1) may spend $O(ln)$ computational complexity, and step (2) may spend $O(lnm)$ computational complexity, since for each query in Q, each candidate competitor and each possible attribute are to be scanned. Clearly, both steps (3) and (4) will spend $O(nm)$ computational complexity, and step (5) spends $O(n^2m)$ computational complexity. Thus, the total computational complexity of the proposed *BGComp* algorithm is $O(lnm + n^2m)$. Usually, compared with the size of preprocessed query log l (e.g., easily tens of thousands or more), the size of focused competitors n (e.g., usually tens or hundreds) and the size of extracted attributes m (e.g., usually hundreds or thousands) are quite small, i.e., $n << l$ and $m << l$. The overall computational complexity is mainly impacted linearly by the size of query log, namely $O(l)$.

To further demonstrate the computational efficiency of the proposed *BGComp* algorithm, scalability experiments on real world query logs were conducted. Figure 3 draws the running time results with respect to different query log sizes given different numbers of focused competitors. The running results explicitly show the linear trend of running times with respect to the sizes of query logs, which is consistent to the theoretical analysis. Besides, with the increase of number of focused competitors, the running times will also increase.

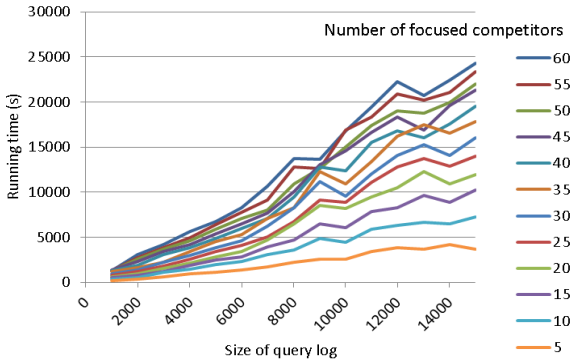

Fig. 3: Results of scalability experiments

5. Sales Forecasting with Competitiveness Degrees

As discussed previously, the big data analysis based on query logs claims that the search volume of querying keywords about a certain product could be used to measure its corresponding market sales. Some work has been conducted on this direction [26, 27, 28]. However, a critically ignored point is that the market is rich and compounded with competitors, revealing a fact that the market performance of a company/product/brand is not only influenced by its attractiveness (e.g., directly reflected by its own query volume), but also interfered by its competitors. This enlightens that the competitiveness degree could also been considered as a key factor in forecasting market sales/shares.

To validate the benefit of competitiveness degrees calculated based on query logs, we conducted an analysis on sales forecasting in U.S. automobile market, since not only it is a typical hyper-competitive market, but some competitive intelligence analyses with UGCs have been conducted on it and show close relationship between the query volumes and the sales of automobiles [25, 26, 27]. Moreover, for comparison purpose, the regression model in [25] is adopted as a benchmark, which was used to forecast automobile sales with query volumes of automobile brands. The simple regression model can serve as the baseline to reveal the effectiveness of the proposed competitiveness degree. Certainly it is possible to build more sophisticated model and this would be improved in future work.

Suppose there are n automobile brands, which can be regarded as n focused competitors with correspond-

ing query volumes, according to [25], the following regression model, i.e., Model (0), is as follows:

$$Sales_{i,t} = \alpha + \beta_1 QV_{i,t-1} + \varepsilon. \quad (5)$$

In Model (0), variable $QV_{i,t-1}$ is the query volume of the brand *i* at period *t*, variable $Sales_{i,t}$ is the actual sales of brand *i* at period *t*. The model is to directly evaluate the forecasting power of query volumes on sales.

Based on Model (0), to further integrate the impact of competitiveness, a new variable $CompQV_{i,t-1}$ is constructed as follows:

$$CompQV_{i,t} = \sum_{j=1}^{n} Comp_{ji} \times QV_{j,t}. \quad (6)$$

Since $Comp_{ji}$, i.e., $Comp(c_j \rightarrow c_i)$, implies to what extent that brand *i* may seize the share of search engine users' intents on the query market from its competitive brand *j*, $CompQV_{i,t}$ represents the weighted query volumes that brand *i* may seize from all of its competitive brands, which can be used as a variable indicating aggregated competitiveness of brand *i*.

With the new $CompQV_{i,t}$ variable, two new regression models could be constructed as follows, called Model (1) and Model (2), respectively:

$$Sales_{i,t} = \alpha + \beta_2 CompQV_{i,t-1} + \varepsilon. \quad (7)$$

$$Sales_{i,t} = \alpha + \beta_1 QV_{i,t-1} + \beta_2 CompQV_{i,t-1} + \varepsilon. \quad (8)$$

Clearly, Model (1) is to directly evaluate the forecasting power of query volumes weighted by competitiveness degrees, while Model (2) is to evaluate the compositional forecasting power of both of them.

In order to analyze the regression models, the data from the well-known automobile website, Edmunds.com, was collected, which is popularly used in competitiveness analysis study [27]. Edmunds.com provides a catalogue of 33 automobile brands, that are popular in the U.S. Market, as well as their monthly sales, from April 2013 to April 2014, which can be used to evaluate the variable $Sales_{i,t}$. The list of the automobile brands on Edmunds.com is as shown in Table 1.

Index	*Brands*	*Index*	*Brands*	*Index*	*Brands*
1	*Acura*	12	*Honda*	23	*Mini*
2	*Audi*	13	*Hyundai*	24	*Mitsubishi*
3	*BMW*	14	*Infiniti*	25	*Nissan*
4	*Buick*	15	*Jaguar*	26	*Porsche*
5	*Cadillac*	16	*Jeep*	27	*RAM*
6	*Chevrolet*	17	*Kia*	28	*Scion*
7	*Chrysler*	18	*Land Rover*	29	*Smart*
8	*Dodge*	19	*Lexus*	30	*Subaru*
9	*Fiat*	20	*Lincoln*	31	*Toyota*
10	*Ford*	21	*Mazda*	32	*Volkswagen*
11	*GMC*	22	*Mercedes*	33	*Volvo*

Table 1: The List of Car Brands on Edmunds.com

It should be noted that, to keep consistent with the sales data in U.S. Market, the query logs were also crawled from Google of U.S. in the same period. It was summarized that the size of the pre-processed query log containing the 33 brands is more than 130,000, and the number of qualified extracted attributes is nearly 2,000. Moreover, to keep consistent with benchmark model, the period unit is set to 1 month, i.e., the regression models are to forecast the next month's sales based on the query volumes and competitiveness degrees of current month.

The calculated regression results are as shown in Table 2 and some valuable implications could be derived. First, all the coefficients of the three models show positive influences on sales at a quite high significance level (i.e., ***), verifying that both search volumes and competitiveness are highly significantly correlated to next period's sales, which echoes existing studies on query log based analysis. Second, the adjusted R^2, representing the forecasting power [30], of the benchmark Model (0) is 50.45%, which is consistent with previous research [25], showing the good forecasting capability hidden in query volumes. Third, the adjusted R^2 of Model (1) is 59.08%, which is 17.11% more than that of Model (0), revealing that weighting competitiveness degrees with query volumes can better help forecast sales, due to the fact that inherent competitiveness in the market has been captured and measured via the proposed method. Fourth, with the assembly of both query volume and competitiveness in Model (2), the forecasting power is much higher, i.e., adjusted R^2 is 65.68% and 30.19% higher than Model (0). As a summary, competitiveness degrees measured based on query logs do be useful in competitiveness intelligence analysis, and can be further utilized to help forecast future market performance, e.g., market sales.

Models	β_1	β_2	*Adj.* R^2	ΔR^2 (%)
Model (0)	1.4795 ***		50.45%	
Model (1)		1.6717 ***	59.08%	17.11%
Model (2)	0.7306 ***	1.1548 ***	65.68%	30.19%

Significance: 0 ***; 0.001 **; 0.01 *; 0.05 .

Table 2: Regression Analysis Results

6. Conclusion

Competitiveness degrees analysis surely is a very important stage in competitive intelligence analysis for companies to seizing market advantages. Query logs on search engine, as a rich source containing massive users' search intents, inherently reflect users' perceptions on competitiveness. However, direct analysis on query logs can hardly discover the competitiveness as well as measuring competitiveness degrees, due to scarce co-occurrence of multiple competitors in the same queries. This paper investigates the joint co-occurrence among competitors via conjunctively queried intermediates, i.e., attributes, and constructs a bipartite graph model accordingly. With the model, the competitiveness relationship can be detected and the mutual competitiveness degrees between any two competitors can be measured. Moreover, a so-called *BGComp* algorithm is designed, which can effectively calculate the mutual competitiveness degrees among competitors. Finally, an analysis with automobiles' sales forecasting is conducted to validate the benefits of the competitiveness degrees calculated by the *BGComp* algorithm.

Future work will center on two directions. On one direction, in-depth investigation will be made to further improve the proposed bipartite graph model as well as the algorithm. The other direction is to carry on more experimental analysis on real world applications.

Acknowledgements

The work was partly supported by the National Natural Science Foundation of China (71372044/71490724/ 71110107027/71402186), and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities of China (12JJD630001).

References

[1] Clark, B.H., Managerial identification of competitors: accuracy and performance consequences. Journal of Strategic Marketing, 2011. 19(3): p. 209-227.

[2] Porter, M.E., Competitive Strategy: Techniques for Analyzing Industries and Competitors. 2008: Free Press.

[3] Zahra, S.A. and S.S. Chaples, Blind Spots in Competitive Analysis. The Academy of Management Executive (1993-2005), 1993. 7(2): p. 7-28.

[4] Elrod, T., et al., Inferring market structure from customer response to competing and complementary products. Marketing Letters, 2002. 13(3): p. 221-232.

[5] Day, G.S., A.D. Shocker and R.K. Srivastava, Customer-Oriented Approaches to Identifying Product-Markets. Journal of Marketing, 1979. 43(4): p. 8-19.

[6] Smith, K.G., W.J. Ferrier and H. Ndofor, Competitive dynamics research: Critique and future directions. Handbook of strategic management, 2001: p. 315-361.

[7] Ketchen, D.J., C.C. Snow and V.L. Hoover, Research on Competitive Dynamics: Recent Accomplishments and Future Challenges. Journal of Management, 2004. 30(6): p. 779 -804.

[8] Lee, T.Y. and E.T. Bradlow, Automated marketing research using online customer reviews. Journal of Marketing Research, 2011. 48(5): p. 881-894.

[9] Ringel, D.M. and B. Skiera. Understanding Competition Using Big Consumer Search Data. in System Sciences (HICSS), 2014 47th Hawaii International Conference on. 2014: IEEE.

[10] DeSarbo, W.S., R. Grewal and J. Wind, Who competes with whom? A demand-based perspective for identifying and representing asymmetric competition. Strategic Management Journal, 2006. 27(2): p. 101-129.

[11] Peteraf, M.A. and M.E. Bergen, Scanning dynamic competitive landscapes: a market-based and resource-based framework. Strategic Management Journal, 2003. 24(10): p. 1027-1041.

[12] Roberts, J.H. and J.M. Lattin, Development and Testing of a Model of Consideration Set Composition. Journal of Marketing Research (JMR), 1991. 28(4).

[13] Siddarth, S., R.E. Bucklin and D.G. Morrison, Making the cut: Modeling and analyzing choice set restriction in scanner panel data. Journal of Marketing Research, 1995: p. 255-266.

[14] Few, W.T., Managerial Competitor Identification: Integrating the Categorization, Economic and Organizational Identity Perspectives. 2007.

[15] Pant, G. and O.R.L. Sheng. Avoiding the Blind Spots: Competitor Identification Using Web Text and Linkage Structure. 2009: Association for Information Systems.

[16] Lappas, T., G. Valkanas and D. Gunopulos. Efficient and domain-invariant competitor mining. in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012. Beijing, China: ACM.

[17] Zheng, Z.E., P. Fader and B. Padmanabhan, From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data. Information Systems Research, 2012. 23(3-part-1): p. 698-720.

[18] Shenghua, B., et al., Competitor Mining with the Web. IEEE Transactions on Knowledge and Data Engineering, 2008. 20(10): p. 1297-1310.

[19] Ma, Z., G. Pant and O.R.L. Sheng, Mining competitor relationships from online news: A network-based approach. Electronic Commerce Research and Applications, 2011. 10(4): p. 418-427.

[20] Xu, K., et al., Mining comparative opinions from customer reviews for Competitive Intelligence. Decision Support Systems, 2011. 50(4): p. 743 - 754.

[21] Kim, J.B., P. Albuquerque and B.J. Bronnenberg, Mapping online consumer search. Journal of Marketing Research, 2011. 48(1): p. 13-27.

[22] Vaughan, L. and E. Romero Fr ás, Web search volume as a predictor of academic fame: An exploration of Google Trends. Journal of the Association for Information Science and Technology, 2014. 65(4): p. 707-720.

[23] Barreira, N., P. Godinho and P. Melo, Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. NETNOMICS: Economic Research and Electronic Networking, 2013. 14(3): p. 129-165.

[24] Da, Z., J. Engelberg and P. Gao, In search of attention. The Journal of Finance, 2011. 66(5): p. 1461-1499.

[25] Choi, H. and H. Varian, Predicting the present with google trends. Economic Record, 2012. 88(s1): p. 2-9.

[26] Bign é, J.E. and N.V. LÓpez, Competitive groups in the automobile industry: a compared supply-demand approach. Journal of Strategic Marketing, 2002. 10(1): p. 21-45.

[27] Ronen Feldman, J.G.A.O., Mine Your Own Business: Market Structure Surveillance Through Text Mining. Marketing Science, 2012. 31(3): p. 521–543.

[28] Carri ère Swallow, Y. and F. Labb é, Nowcasting with Google Trends in an emerging market. Journal of Forecasting, 2013. 32(4): p. 289-298.

[29] Wu, L. and E. Brynjolfsson, The future of prediction: How Google searches foreshadow housing prices and sales, in Economics of Digitization. 2013, University of Chicago Press.

[30] Draper, N. R., Smith, H., and Pownell, E., Applied regression analysis. 1966, New York: Wiley.