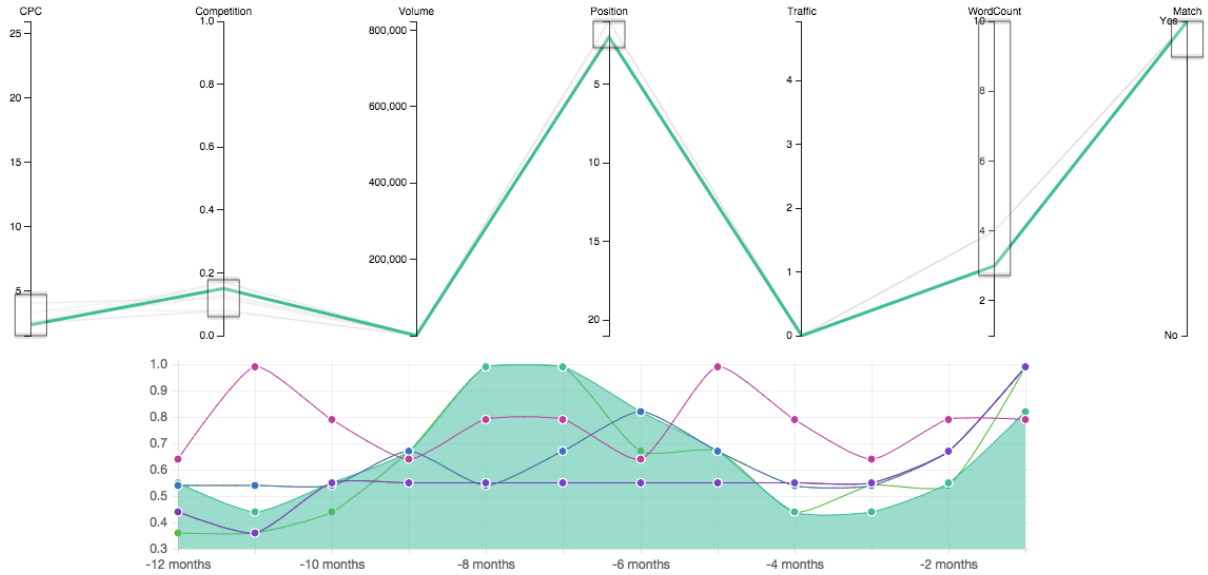
KeywordVis: Identifying and Analyzing Keywords for Search Engine Advertising

Yingsai Dong
University of British Columbia



Domain: hotels.com

#Selected Keywords(max 15): 1 Click the row to compare the keyword. Counter(max 7): 7 Hide Trends

Showing all 12 rows

Keyword	CPC	Competition	Volume	Position	Traffic	WordCount	Match	select
hotels near the airport	2.49	0.08	480	1	0	4	Yes	In List
hotels by the airport	2.25	0.17	480	1	0	4	Yes	+
hotels near us	2.57	0.13	260	1	0	3	Yes	+
hotels in area	**2.36**	**0.15**	**590**	**2**	**0**	**3**	**Yes**	**+**
hotels by the airport	2.25	0.17	480	2	0	4	Yes	+
hotels near the airport	2.49	0.08	480	2	0	4	Yes	+
hotel customer service	4.03	0.12	110	1	0	3	Yes	+
hotels near us	2.57	0.13	260	2	0	3	Yes	+
hotels on the ocean	2.56	0.17	50	1	0	4	Yes	+

Fig. 1. With KeywordVis, users can compare ads keywords marked in different colours with linked highlighting in a parallel coordinates panel and a search volume trends line chart.

Abstract—Medium-small advertisers need to find profitable keywords when they set up search engine advertising campaign for their domains. Most of the existing ads keyword research tools are either visualizing keywords with inadequate data or just showing a big table of numbers without visualization, thus users can not effectively discover the keywords which can reach their target users. We design and implement KeywordVis, an online visualization application, to utilize data from SEMrush.com to help medium-small advertisers with four tasks: 1) deriving new attributes to indicate relevance, 2) efficiently determining a list of profitable keywords with user specified filters, 3) comparing attribute values across keywords and 4) comparing advertiser's own list with the competitor's in an comprehensive analysis view. KeywordVis is carefully designed with easy-to-use interface and sufficient functionalities to fully support the four tasks.

Index Terms—Search Engine Advertising, Keyword Research, Information Visualization

1 Introduction and Background

Online market has become a significant business platform for people who sell product or provide services. Huge amount of online transactions are taking places every second. Searching online is foremost among the approaches one takes for shopping planning. As a consequence, online advertising is becoming the most efficient way to make people aware of websites and promote offers. Online advertising can be categorized into two branches: search engine ads and display ads. The former can propel one's business onto the front page of search results, while the latter displays ads when people are having everyday

• *E-mail: yingsaid@cs.ubc.ca*

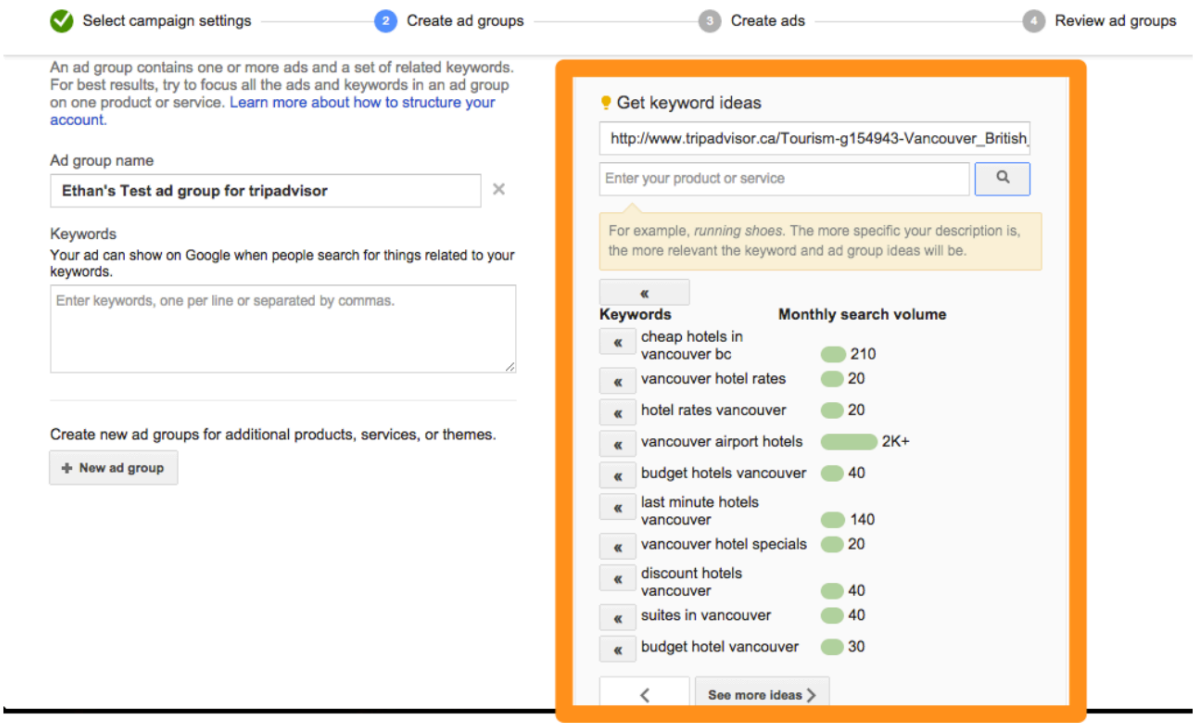


Fig. 2. Keywords suggestion using AdWords when starting a new ads campaign

activities, for example, watching videos, browsing articles, checking emails and so on. Among these two approaches, medium-small advertisers would prefer the search engine ads because it is more dedicated to people with direct and immediate purchasing desire. On the other hand, display ads are more helpful for big advertisers with sufficient budget who want to retain their impact among customers. As a downside, displaying ads can sometimes interrupt one's online activity and become arguably more annoying than search engine ads (Much of the mechanism of display ads is beyond the scope of this work, we would make no further discussions about it beyond this comment).

There are several different advertising platforms provided by those big search engine companies: Google, Microsoft, Yahoo! and so on. We will focus on the AdWords platform of Google. Google AdWords is the best-known platform of keyword advertising with a dominant market share. Google AdWords also has a dominant number of users working on their ads campaign in a daily base. Thus, AdWords is able to provide a large database with a high quality ecosystem. Basically, the top four important things to set up ads campaign using AdWords are:

A. **Bidding.** Almost all the advertisers using AdWords leave the job of setting up a bidding strategy to Google. It would take them much more effort to come up with a good bidding strategy. Besides, even if they can design their own bidding strategies, it might not be better than the default ones Google provides.

B. **Budget.** You get to choose a daily budget. Once this budget is reached, Google will stop showing your ads until the next day. The budget won't actually affect much on your ads campaign performance. What matters are the ads content and keywords you choose as we will introduce later. The budget determines how long your ads campaign will run every day. On the other hand, ads content and the keywords we choose determines the rate of return as a measure of how well our ads performs. Budget is used to calculate the actual amount of profit by multiplying itself with the rate of return. One naive idea to set up keywords could be to simply select all relevant ones and eliminate bad ones by looking at the long-term performance. For advertisers with limited marketing budget, this idea could be too expansive to be carried out, while for those with sufficient budget, this idea is obviously too time consuming. Both of them need to choose keywords strategically to maximize potential impact while keeping cost (both in money and time) as low as possible.

C. **Ads Description.** There are actually a lot of tips on how to write ads titles and content. Usually, advertisers want to make them as specific to your product as possible and emphasize the benefits.

D. **Choosing keywords.** Choosing the right keywords is vital to ads campaign: choose the right keywords, and your campaign will bring in new customers; pick the wrong keywords, and you will either be completely ignored or –worse– your campaign will be an expensive flop. This is actually where advertisers should put most of their efforts on. When you start a new ads campaign, you have to learn the language the users use when they are looking for the product and service you offer on your website. For example, you think you are running a "pre-owned car" business, however you are actually in the "used car" business, simply because much more people put the keyword "used car" instead of "pre-owned car" when they search on Google. AdWords provides very limited service on choosing keywords. The example of starting a new ads campaign using AdWords is shown in figure 2. In this example, I am starting a new campaign to advertise the Vancouver page of the tripadvisor.com. When it comes to the step of choosing keywords, it will show you a list of related keywords based on the URL you provide. As you can see, the information you have to help choosing keywords from the list is extremely limited. There is only a rough number of monthly search volume which is far from enough to let you make good choices. You will only have a good evaluation of the keywords when there are more associated attributes such as Cost-Per-Click (CPC), competition level, and the position your website appears when searching this keyword. A lot of people would just come up with keywords by brainstorming or even randomly choose some potentially good keywords. Then they would just start with these keywords and wait to check how the different keywords perform in order to

eliminate the bad ones. There are critical downsides: some profitable keywords may not be brought in your original list and they would hardly be caught later; this process could probably take significant amount of time to stabilize your keywords strategy and by then you would have wasted a lot of money. Therefore, people created some tools helping you choose the right keywords based on more detailed data. We will discuss those tools in the following section.

Among the above four components, choosing keywords is the important one. Google Keywords Planner(GKP) and SEMrush are the best two ads keyword research tools. They both give you a long list of keywords with some associated attributes for a given domain. The long list of keywords is also called an **organic keywords report**, where "organic" implies that the website domain appears among the non-ads search results for the given keywords, not as one of the ads. GKP gets the data from Google itself, which guarantees the data accuracy. At the same time, SEMrush is a Google partner, also providing organic report for a given domain. SEMrush only analyzes Google's top 20 results because this is where over 99% of all search engine traffic comes from. Therefore, in terms of data accuracy, people consider it a tie. However, what makes them different is that GKP only gives you part of the attributes SEMrush provides, namely the average monthly searches, suggested bids, and a categorical competition level. Therefore, all the data we use for KeywordVis comes from SEMrush.com.

On SEMrush, if you don't pay, you could only get the first 10 lines of data for one keywords report. If you have a 69$ per month subscription, you can get at most 10,000 lines for a report, which is what we chose. A typical approach medium-small advertisers take is to run their online businesses for a while without ads campaign running. During that period of time, some initial search data would be generated and can be stored by SEMrush. Any keyword people search on Google with a domain listed in the top 20 results will be in the report for that domain.

We designed KeywordVis, an online visualization application, to utilize the data from SEMrush.com to help the medium-small advertisers choosing right keywords when starting a new ads campaign.

2 Related Work

We introduce the previous work on visualizing ads keywords and the related work on the design idiom we use in KeywordVis.

2.1 Visualizing Ads Keywords

Although a lot of websites provide ads keywords data, there are only limited previous work on visualization of the data. One of the few ads keywords visualization tools is Keyword Eye [8] which is the most popular one. Keyword Eye has been online from 2010. It utilizes the data from SEMrush.com as well. Advertisers can use Keyword Eye to visualize an organic report for a domain.

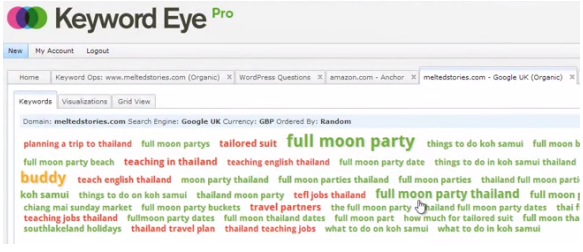


Fig. 3. Visualization of organic keyword report using Keyword Eye.

An example for meltedstories.com is shown in figure 3. All the keywords in the report are encoded as a "keyword cloud" ordered by any specified attribute. The size of the keywords is proportional to the search volume value. The three different colours indicate the three competition levels: high, medium, and low. The second tab visualizes the search volume share and correlation between search volume

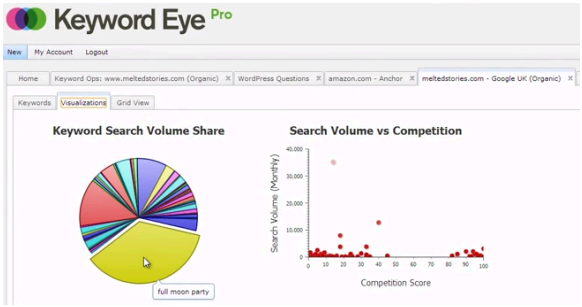


Fig. 4. Visualization of Volume share and using Keyword Eye.

and competition as shown in figure 4. The last tab shows the table of keywords data.

There are some limitation if advertisers want to effectively find profitable keywords using Keyword Eye. First, besides search volume, competition, and the attribute for ordering, some other important attributes are not visually encoded in the "keyword cloud". Second, it is impossible to filter the keywords on multiple attributes. Third, the font size and three colours are very inaccurate to indicate the values of search volume and competition. If users want to refer to the exact values of all the attributes, they have to switch to the last tab to manually find the keywords in the table. Fourth, the pie chart and scatterplot are independent from each other. For example, hovering over a keyword in the pie chart doesn't highlight the same keyword in the scatterplot. Fifth, the only one thing users can do for selected keywords is to download them. Users can not analyze the list of selected keywords and compare it with another list. KeywordVis can eliminate all the above limitation and allow users to efficiently find and analyze keywords to set up ads campaign for their domain.

2.2 Parallel Coordinates

The most important visualization idiom we use in KeywordVis is parallel coordinates. Parallel Coordinates was initially presented by Alfred Inselberg in 1985 [5]. It was presented to solve the problem that the perceptual experience of higher-dimensional spaces is limited by our 3-dimensional habitation. People were not satisfied to explore properties of such geometries only in the abstract. Since then, parallel coordinates has been widely used as a visualization technique for multivariate data and high-dimensional geometry [4]. Parallel coordinates are mostly used to model and detect correlation [6]. After more interactive dimension management (for example, dimension ordering, spacing, and filtering) were enabled [10], parallel coordinates became more popular as a part of visualization tool to effectively explore the multi-dimensional dataset. In addition, more front-end techniques have been implemented to make it easier for developers to utilize parallel coordinates in different tasks, for example, brushing on axes [3].

To our best knowledge, KeywordVis is the first tool to visualize ads keywords data using parallel coordinates. Based on the research work of Yang et al. [10] and Rosenbaum et al. [9], we enable dimension re-ordering and progressive rendering to better achieve the goals of KeywordVis. We will introduce the design of KeywordVis with more details later.

3 Data and Task Abstraction

3.1 Data Abstraction

When you give SEMrush a domain name, it will show you a full keywords report. For example, the organic keywords report for autotrader.com, a car trading website, is shown in figure 5.

The dataset is a table. Every row in this table represents a keyword. Every column indicates an attribute associated to the keyword. As I mentioned before, the size of this table could be as large as 10,000 rows. We could export this table as a CSV file. This is the data source

ORGANIC SEARCH POSITIONS 1 - 100 (444,002)

Filter by keyword | Filters | Export

Keyword	Pos	Volume	CPC	URL	Traffic %	Costs %	Com.	Results	Trend	SERP source	Last update
autotrader	1 (1)	5,000,000	0.12	www.autotrader.com/	27.32	1.48	0.13	21,700,000			3 hr ago
auto trader	1 (1)	673,000	0.22	www.autotrader.com/	3.67	0.36	0.12	48,700,000			3 hr ago
used cars	1 (1)	450,000	2.46	www.autotrader.c...-cars/	2.45	2.73	0.91	337,000,000			3 hr ago
used cars for sale	1 (1)	301,000	2.80	www.autotrader.com/	1.64	2.08	0.92	198,000,000			3 hr ago
autotrader com	1 (1)	201,000	0.14	www.autotrader.com/	1.09	0.06	0.20	8,650,000			3 hr ago
cars for sale	3 (3)	368,000	2.47	www.autotrader.c...le.jsp	0.38	0.42	0.90	456,000,000			3 hr ago
car dealerships	1 (1)	60,500	6.45	www.autotrader.com/car-dealers	0.33	0.96	0.90	39,000,000			3 hr ago
used car dealerships	1 (1)	49,500	4.71	www.autotrader.com/car-dealers	0.27	0.57	0.94	18,700,000			3 hr ago
cars for sale by owner	1 (1)	49,500	1.35	www.autotrader.c...er.jsp	0.27	0.16	0.87	159,000,000			3 hr ago
www autotrader com	1 (1)	49,500	0.13	www.autotrader.com/	0.27	0.01	0.18	9,780,000			3 hr ago
car trader	1 (1)	40,500	0.93	www.autotrader.com/	0.22	0.09	0.56	86,200,000			3 hr ago

Fig. 5. The organic keywords report for autotrader.com

we used for this project. All the attributes in an organic keywords report are analyzed in details as follows:

- **Keyword.** The keyword phrase bringing users to the website via Google's top 20 organic search results. The expression "keyword phrase" will be used when necessary to precisely indicate the keyword string to avoid misunderstanding, otherwise just "keyword". It is a string in the exported file.

- **Pos.** The position the domain gets in the search for the given keyword in the current month. The number in bracket shows the previous position in last month. It is an integer number in the exported file.

- **Volume.** The average number of search queries for the given keyword in the last 12 months. It is an integer number in the exported file.

- **CPC.** Average price in U.S. dollars advertisers pay for a users click on an ad for the given keyword using Google AdWords. It is a floating number with two digits after the decimal point in the exported file.

- **URL.** The URL displayed in search results for the given keyword. It is a string in the exported file.

- **Traffic%.** The share of traffic driven to the website with the given keyword in the current month. It is a floating number with two digits after the decimal point in the exported file.

- **Costs%.** The estimated proportion of the estimated cost of buying the same number of visitors for a term, as compared to the estimated cost of the same number of targeted visitors coming to this site organically. It is a floating number with two digits after the decimal point in the exported file.

- **Com.** Competitive density of the advertisers using the given keywords for their ads. One(1) means the highest competition. It is a floating number with two digits after the decimal point in the exported file.

- **Result.** The number of URLs displays in organic search results for the given keyword. It is an integer number in the exported file.

- **Trend.** A list of 12 quantitative numbers representing the interest of searches in the given keyword during the last 12 months. The metric is based on changes in the number of queries per month. These 12 numbers are normalized to a range between 0 and 1. They are 12 floating numbers with two digits after the decimal point indicating this trend in the exported file.

- **SERP source.** A snapshot of the search engine result page (SERP) for the given keyword. It is not contained in the exported file.

- **Last Update.** The time when the given keyword was last updated. It is an integer number indicating a time stamp in the exported file.

In terms of domain dependency, among these 11 attributes, the values of Pos, URL, Traffic%, Costs%, and SERP source for a keyword depend on the domains. The values of all the other 6 attributes for a given keyword are independent across domains. However, when it comes to choosing the useful attributes for evaluating the ads keywords, we would take this group of attributes: Position, Volume, CPC, traffic%, Com., and Trend, as a mix of dependent and independent attributes. URL, Results, and Last update are barely useful in evaluating keywords. SERP source is not provided in the exported file. Cost% is a highly unreliable estimation SEMrush provides and advertisers don't use it as a useful parameter when choosing keywords.

In terms of attribute types, almost all the attributes are quantitative except Pos, URL, and SERP source. Pos is actually a ranking number the domain gets in the first 20 results from Google. It ranges from 1 to 20. Volume, CPC, and Results data come from Google. SEMrush does not provide a explicit maximum value for them. The ranges for both Traffic% and Costs% are 0 to 100. Com. data ranges from 0.00 to 1.00. Trends data ranges from 0.00 to 0.99. All the ordered data among these attributes are in sequential direction of ordering.

A comprehensive review of attribute type, domain dependency, and whether used in in this project is shown in figure 6. To be more specific, there is one such table per domain. Thus, the dataset we would

actually use in this project is a N $\times$ 7 table ($N \leq 10,000$) for one domain. In the following section, when it comes to the task of comparing two domains, two such tables will be used as our data source.

Attributes	Attribute Type	Independent from domain	Used in this project
1. Pos	Ordinal (Range:1-20)	No	Yes
2. Volume	Quantitative (Range: >=0)	Yes	Yes
3. CPC	Quantitative (Range: >=0)	Yes	Yes
4. URL	String	No	No
5. Traffic%	Quantitative (Range: 0-100)	No	Yes
6. Costs%	Quantitative (Range: 0-100)	No	No
7. Com.	Quantitative (Range: 0.00-1.00)	Yes	Yes
8. Results	Quantitative (Range: >=0)	Yes	No
9. Trends	12 Quantitative Numbers (Range: 0.00-0.99)	Yes	Yes
10. SERP source	Snapshot	No	No
11. Last Update	Quantitative (Time stamp)	Yes	No

Fig. 6. Analysis of the attributes.

There is one more thing worth mentioning here. SEMrush also provides a useful feature that users could get a table of all the keywords a domain has used for the ads campaigns. Different from the organic report in figure 5, it is called a **paid keyword report**. Everything is the same expect one attribute: Position. In the paid report, the position attribute indicates the position of the ads for the domain among all the ads shown besides the search results, instead of the position of the domain among all the non-ads searching results. The review of the position attribute in the paid report is shown in figure 7. The max number of ads shown in the front page is 11, which explains the range.

Attributes	Attribute Type	Independent from domain	Used in this project
Pos (in paid report)	Ordinal (Range:1-11)	No	Yes

Fig. 7. Analysis of the position attribute in the paid report.

3.2 Task Abstraction

3.2.1 Domain-specific Tasks

The three most important features people care about when evaluating the profitability of a keyword:

1. **Search Volume.** How often people type this keyword when searching. Choosing keywords is a balance between cost and search volume. Expensive keywords are likely to have a bigger volume, because more people search for them. But they will also burn through your ads budget quickly. Lower priced keywords have a smaller volume, so they may not bring in the level of traffic you need.

2. **Competition.** You are always competing against others in an auction for the keywords you are targeting for your campaign. The more competitive a keyword is, the higher CPC it has.

3. **Relevance.** Advertisers always want those people searching the keyword are likely to be the same group of people who will click on their ads and then take an desired action on the landing page, for instance filling out a form, making a purchase or signing up. For example, a person searching for "shoes" is probably browsing, and not ready to buy. On the other hand, someone searching for "best price on Air Jordan size 12" practically has his wallet out! Typically a more specific keyword means more relevance to the product and service offered by your website.

The data we have from SEMrush directly have corresponding metrics for the Search Volume and Competition. But there is no direct data indicating relevance. The relevance is not as simple as the other two features. It could hardly be represented using just one attribute. One of the goals of KeywordVis is to provide as much attributes indicating relevance as possible, so that users would have more control on how to evaluate and discover the keywords they want.

Based on the three features mentioned above, the **long-tail keywords** are exactly the profitable keywords medium-small advertisers want to find. Long-tail keywords are the ones that may not get a lot of searches and are not in high competition, but they are extremely specific and extremely likely to convert. And if we add them all up, they actually count for a lot of profitable clicks. A list of long-tail keywords is what the advertisers could get as one of the outputs of this vis system. We will use the expression **"A-list"** to represent this output list of keywords advertisers finally choose for the domain. When the advertisers are analyzing keywords and determining the A-list, comparisons on different attributes across keywords should also be enabled.

Another big challenge for medium-small advertisers are that they are usually facing some competitors. As I mentioned in the end of section 3.1, if a domain has already run ads campaign, SEMrush can give you a report showing what keywords the domain has used. When you provide the domain name to SEMrush, it will show you not only the organic report but also a paid report. The attributes in the paid report are almost exactly the same as those in the organic report except the Position attribute. We found that even for small-medium advertisers, the number of keywords in the paid report is usually larger than 15, for example, 50 to 100, or even around 1,000. This is actually because this paid report contains all the keywords ever used in the past 12 months, given that the Volume attribute is an average of the last 12 months. However, the Traffic% attribute can help us find the subset that is currently active because Traffic% indicates the share of traffic in the current month. We observed that if you order the keywords in an descending order on Traffic%, the value of Traffic% typically falls sharply to almost zero within the top 15 keywords. In another word, the currently active subset of keywords are typically among the top 15 based on Traffic%. Thus, we could reduce the list of keywords to this subset. We will use the expression **"C-list"** to represent this reduced list of keywords a certain competitor's domain is currently using. It is safe to set 15 as an upper limit of the C-list size. Advertisers would be very interested in analyzing the difference between A-list and competitor's C-list.

3.2.2 Tasks in Abstract Forms

A. **Derive new attributes.** To provide more attributes for keyword relevance, we will derive these two new attributes:

   - **Word Count.** Word count of the keyword phrase is one way to determine how specific the keyword is. The more words a keyword phrase has, the more specific it is. This new attribute allows users to transform the keyword phrase from a descriptive string to an integer number. The attribute type is quantitative.
   - **String Match.** Whether a user-specified string is contained in the keyword phrase. One common strategy is that the advertisers would like to use the keyword phrases containing strings exactly matching the title or the critical words in the ads description. This vis system will enable users to manually specify the string. The attribute type is boolean.

   When the list of keywords for one domain is initialized in our vis system, Word Count attribute will simultaneously be initialized. If users specify a string, we will have String Match attribute generated. These two derived attributes extend the dataset beyond the original set of the 6 attributes. We would have four attributes for relevance: Position, Traffic%, Word Count, and String Match.

B. **Determine a list of keywords.** Advertisers can use this vis system to determine the A-list containing up to 15 keywords, presumably a list of long-tail keywords with high relevance to their

product and service, search volume lower than a certain value, and Competition also smaller than a certain value. The advertisers will take the A-list to set their new ads campaign. Typically advertisers want to have a list of 5-15 closely related keywords. This size gives them the ability to scale their campaigns when organizing, while still being able to serve relevant copies of ads description to the search users. Setting size no more than 15 is also widely recommended in most of the AdWords tutorials. Thus, it is safe to set 15 as a upper limit for the A-list size.

C. **Compare attribute values across keywords within a list.** Comparing keywords against each other is a common practice when advertisers are doing keywords analysis. In order to analyze how differently keywords are performing and uncover the different patterns on search volume trends, advertisers can put some keywords in a so-called **compare-list** to compare the the attribute values across keywords within this compare-list.

D. **Compare two lists of keywords.** As mentioned in the end of section 3.1, advertisers are very interested in comparing the A-list with the C-list. We should just allow advertisers to make pairwise comparisons between the A-list and C-list of one competitor, instead of multi-list comparisons. Notice it is more intuitive and effective to compare two lists based on the following two sub-tasks:

- Identify the keywords appearing in both two lists.
- Compare the average value and distribution on five quantitative attributes: Word Count, CPC, Volume, Competition, and Traffic%.

Among the attributes used in this project, position is not a comparable attribute when comparing A-list and C-list since position attribute in the organic report and paid report have different semantics. Comparison between A-list and C-list could effectively tell the difference of keywords investing strategies and overlapping area of the products and services advertiser share with your competitor.

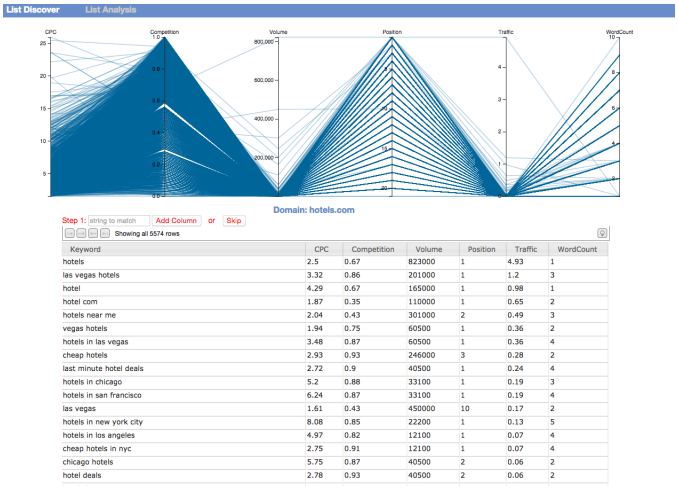
4 Solution

This visualization system is designed to help medium-small advertisers to discover the right keywords to start their new ads campaign and analyze the keyword list. Specifically, to support the four tasks A to D mentioned in section 3.2.2, we designed two views in this visualization system: List Discover and List Analysis. Tasks A, B, and C can be conducted in the List Discover view, while task D in List Analysis view.

4.1 List Discover View

By default, users will see two main components, a parallel coordinates panel on top and a table at the bottom, as shown in figure 8. The design idiom of parallel coordinates is an approach especially for visualizing multiple attributes at once using spatial position. Parallel coordinates visually encode data with two dimensions of spatial position. There are many vertical axes, each for a specific quantitative attribute. Each keyword is represented by a blue line across the axes. In the very beginning users will see six axes corresponding to the three features mentioned in section 3.2.1. From left to right, they are CPC and Competition for Competition feature, Volume for Search Volume feature, Position, Traffic, and Word Count for Relevance feature. The max and min boundaries of the axes are calculated as the max and min values of the data. For the position axis, the smallest value 1 is on the top. For all the other 5 axes, the max values are on the top. In this way, lines crossing the three axes for relevance feature at higher points all means keywords with higher relevance level.

Another alternative people tool usually use for visualizing multiple attributes and performing multidimensional filtering is the Crossfilter JavaScript library. The first advantage of parallel coordinates design over the Crossfilter design is that the parallel coordinates design occupies less display space. this advantage becomes more obvious with



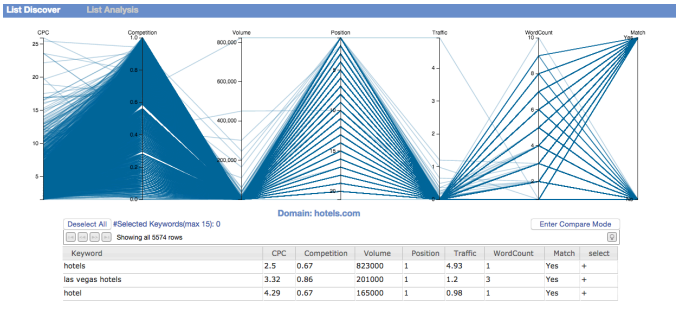
Keyword	CPC	Competition	Volume	Position	Traffic	WordCount
hotels	2.5	0.67	823000	1	4.93	1
las vegas hotels	3.32	0.86	201000	1	1.2	3
hotel	4.29	0.67	165000	1	0.98	1
hotel com	1.87	0.35	110000	1	0.65	2
hotels near me	2.04	0.43	301000	2	0.49	3
vegas hotels	1.94	0.75	60500	1	0.36	2
hotels in las vegas	3.48	0.87	60500	1	0.36	4
cheap hotels	2.93	0.93	246000	3	0.28	2
last minute hotel deals	2.72	0.9	40500	1	0.24	4
hotels in chicago	5.2	0.88	33100	1	0.19	3
hotels in san francisco	6.24	0.87	33100	1	0.19	4
las vegas	1.61	0.43	450000	10	0.17	2
hotels in new york city	8.08	0.85	22200	1	0.13	5
hotels in los angeles	4.97	0.82	12100	1	0.07	4
cheap hotels in nyc	2.75	0.91	12100	1	0.07	4
chicago hotels	5.75	0.87	40500	2	0.06	2
hotel deals	2.78	0.93	40500	2	0.06	2

Fig. 8. Default layout of the List Discover view.

increasing number of attributes, for example 6 or 7 in this project. Second, the data are visualized only in aggregation level in the Crossfilter view, whereas individual items are visualized in the parallel coordinates view. With KeywordVis, when users are determining the keyword list, they can be more focused on the individual keywords than how the aggregations are distributed. Third, in the Crossfilter view, users lose the information of the values on different attribute for a given keyword, while the values are clearly visualized in the parallel coordinates panel of KeywordVis. Fourth, when users are conducting the task C to compare keywords within a list, the comparison can't be visualized in the Crossfilter view, while one can effectively do this with KeywordVis as we will introduce later.

The table in figure 8 shows the keywords data that are visualized in the parallel coordinates panel. Each column of the table by its order relates to a corresponding attribute axes in the parallel coordinate panel. By default the keywords in the table are in the descending order on Traffic attribute. However, you can always re-order the keywords in ascending or descending order by any attribute. We associate the parallel coordinate panel with a table to alleviate the difficulties a user would encounter to understand and interact with the panel alone.

All the keywords data initially shown in the list discover view is from the organic report on SEMrush for a specific domain. The upper limit is ten thousand. In the following three subsections we will introduce the design details from the aspect of tasks.



Keyword	CPC	Competition	Volume	Position	Traffic	WordCount	Match	select
hotels	2.5	0.67	823000	1	4.93	1	Yes	+
las vegas hotels	3.32	0.86	201000	1	1.2	3	Yes	+
hotel	4.29	0.67	165000	1	0.98	1	Yes	+

Fig. 9. Layout of the List Discover view.

4.1.1 Task A: Derive Attributes

The first derived attribute Word Count is automatically generated in the very beginning. It is a quantitative attribute indicating how many words are contained in the keyword phrase. As you can see, the first decision you should make is to specify a string to match. After type the string and click "Add Column", a new column is added to the table

and a new axis is added to the parallel coordinates panel accordingly. It is a boolean attribute with value of “Yes” or “No”. The example of setting the string as “hotel” is shown in figure 9. You can also skip this step.

4.1.2 Task B: Determine a list of keywords

After the first step, users can start working on determining a list of keywords. This operation is indicated by an additional ”select” column at the rightmost of the table with a plus sign in each row. Clicking the plus sign can put the keyword into the current list of selected keywords. Users can find the long-tail keywords by setting specific value ranges on different attributes. To do this in the List Discover view, they can brush along a specific axis to specify a range to filter keywords out as shown in figure 10. The filtered out keywords are also removed from the table. Double-clicking the axis title can change the direction of the axis. Users can also re-order the axes by dragging the axis title in case they are interested in the correlation of some axes. With the String March attribute, we could easily filter out keywords with or without a string by dragging along the axis just like how we do for other attributes. Users can also single-click the axis to remove the range restriction. All the columns are sortable in the table. This is useful when the keywords in your specific range on an attribute are too dense to be distinguished from each other and there are still more than what you want. One can re-order the keywords on the corresponding attribute and choose them from the table.

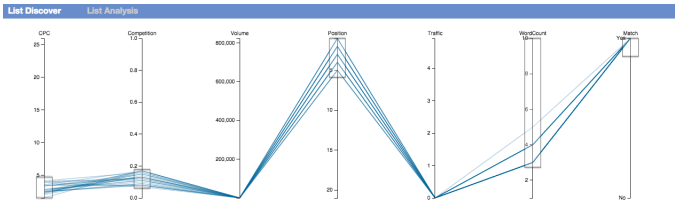


Fig. 10. Filtering by brushing and dragging.

While the users are tuning the ranges, they are narrowing down the keywords candidates. Whenever they think a specific keyword should be selected into the final list, they could click the table cell with plus sign. Then the corresponding row in the table and the corresponding line in the parallel coordinates panel will both be highlighted in red. The plus sign will change to “In List”, and users can deselect the keyword by clicking this table cell. If users still want to select more than 15 keywords, they will not be allowed to do so and will see a notification for the size limitation.

Whenever the list table is not empty, the “Save Selected Keywords” button will present as shown in figure 11. Users can save a list anytime they want by clicking the button and typing a list name. They can save different keywords lists as many as possible. All the lists they have saved can be reviewed and analyzed in the List Analysis view. We will introduce that later in section 4.2.

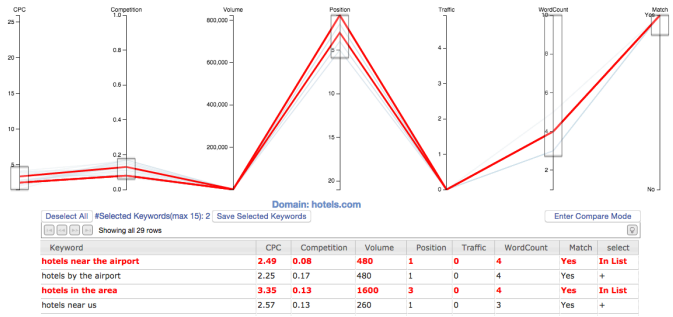


Fig. 11. Selected keywords highlighting and save button.

4.1.3 Task C: Compare keywords within a list

Comparing attribute values across keywords of interest is a common practice for keywords analysis. We allow users to conduct this task in the list discover view. At any moment during the process of selecting keywords in the list discover view, users can enter the compare mode by clicking the “Enter Compare Mode” button. All progress so far will be kept as is. However, all the lines in the parallel coordinates panel are coloured grey to visually notify the users they are in compare mode. All the rows of selected keywords are coloured back to black. Users can add one keyword for comparison by clicking anywhere in the row. The row and the line in parallel coordinates panel will keep highlighted in a unique new colour. Clicking the plus sign cell and clicking the whole row are used to differentiate the tasks of adding keywords to A-list and compare-list on purpose. The upper limit of compare-list size is 7. In this way, there will be up to 7 different colours used here. The colours are selected evenly along the colour hue wheel. There are two reasons to set the upper limit as 7. First, based on a number of experimental uses of this vis system in different scenarios, when the users come to compare mode, it is usually the case that the compared keywords have very similar or even exactly the same values in 3 or more attributes given all the restrictions users put on the axes. The lines are overlapped in a limited area and can only differentiate each other in the rest one or two axes. If there are more than 7 keywords in the compare-list, it can be much harder to recognize the individual lines in the parallel coordinates panel. Second, if more than 7 colours are assigned for each keywords, no matter how they are selected, some of the colours can be harder for users to differentiate.

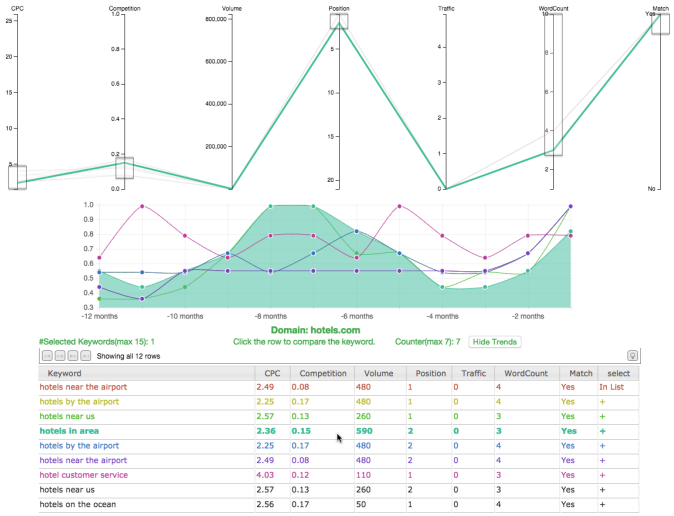


Fig. 12. Highlighting compared keywords in compare mode.

While the compare-list is not empty, there is a “Show Trends” button presenting. If users want to compare the keywords in terms of the search volume trends in the last twelve months, they can click this button and a line charts will appear between the parallel coordinates panel and the table. The reason of choosing line charts over other idioms is that trends pattern can be best emphasized by the connecting lines in line charts. This is the best situation to use the trends data we keep from the organic report. Hovering over the compared keywords in the table can highlight both lines in the parallel coordinates panel and trend lines in the line charts with shared colours shown in figure 12. This creates the linkage across the three components in the view. In the compare mode, users can also add and remove selected keywords for A-list based on the comparison. When users are in the compare mode, the text on “ Enter Compare Mode” button becomes “Exit Compare Mode”. Users can click it to return back to the regular mode.

4.2 List Analysis View

In the List Analysis view, there are two drop-down menus, the left one for selecting lists saved by users, the right one for selecting the pre-saved competitor keyword lists. Every time users save one list in List Discover view, a corresponding option is automatically added to the left drop-down menu here. By default, the first saved list by users and the first competitor list are selected.

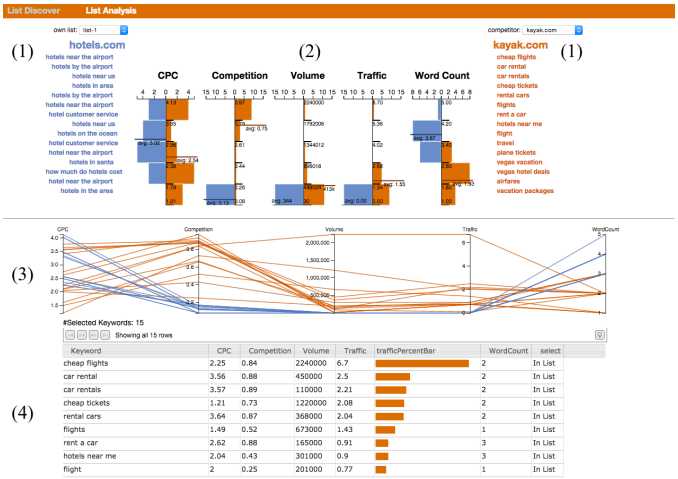


Fig. 13. Layout of the List Analysis view.

Fig. 14. Marking the shared keywords.

4.2.1 Task D: Compare two lists

When selecting two lists to compare as shown in figure 13, there are four components in the List Analysis view: (1). two lists of keywords on each side; (2). bar charts on top in the middle; (3). a parallel coordinates panel for reference; (4). a table showing the keywords in the competitor list. Once users select different list in either drop-down menu, all the data is updated immediately. Across all the colour-marked elements in the the List Analysis view, blue represents keywords from user-saved list, while orange represents keywords from competitor list. The designing of the four components is:

1. **Two lists of keywords on the side.** If there are keywords shared by both lists, they are marked within a pair of chevrons in the lists of keywords on each side as shown in figure 14.

2. **Bar charts on top.** Bar charts are visually encoded as five pairs of vertical histograms with the data of two lists on left and right sides respectively. This design enables users to compare the two lists in terms of the distribution and average values on the five comparable quantitative attributes of the keywords intuitively. This is also where you can effectively check how the keywords

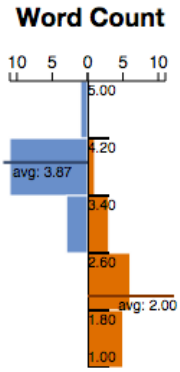


Fig. 15. A pair of vertical histogram for Word Count attribute.

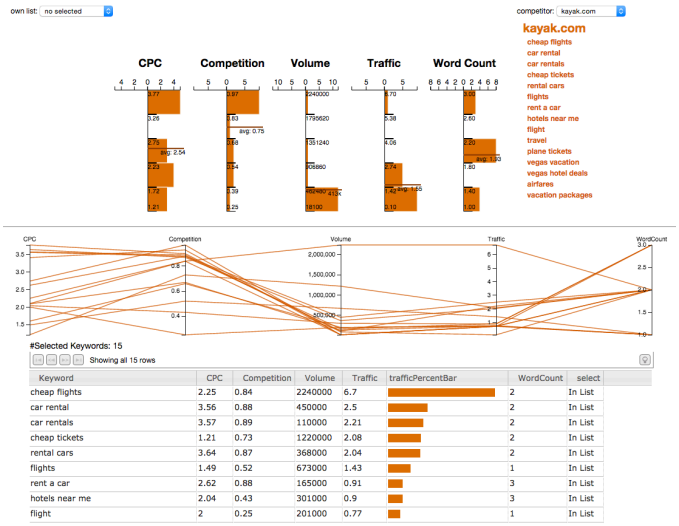


Fig. 16. Layout of the List Analysis view.

you selected satisfy the requirements of long-tail keywords. An example of the pair for Word Count attribute is shown in figure 15. If you hover over a bar, the corresponding keywords represented by this bar are highlighted both in the list on side and in the parallel coordinates for reference.

3. **Parallel coordinates panel for reference.** The parallel coordinates panel always shows the lines representing keywords from both lists.

4. **The table for competitor list.** The table at the bottom shows all keywords from competitor list based on the data from the paid report for the given competitor domain. By default, the keywords are ordered in descending order on Traffic attribute. As mentioned in the end of section 3.2.1, we only keep the top 15 keywords with most traffic to initialize the competitor list. One additional column is created to show bars of traffic data. The length of the bars equals the value of the traffic. The bars can help users easily observing the sharp falling of the traffic within the top 15 keywords. In this way, users can decide whether they should remove the inactive keywords with extremely little traffic from the competitor list. They can do this by clicking the cells in select column. If users changed the competitor list, an "Update Competitor List" button appears. Whenever users think they are done updating the competitor list, they can click this button to update all the other components.

4.2.2 List analysis without list-to-list comparison

If users just want to review the analysis of their own lists or competitor list without list-to-list comparison, they could simply select the empty option from either drop-down menu. All the components are updated with keywords from only one list. An example of analyzing only the competitor list is shown in figure 16.

5 Implementation

This visualization system is implemented as a web application with the current web standards model: HTML, JavaScript and CSS. I have used the following Third-Party JavaScript libraries/toolkits:

- **D3.js.** D3.js is a library for interacting with HTML documents based on data [1]. I used D3.js to load the external data from CSV file. Instead of manipulating the documents with the tedious methods from the standard W3C DOM API, I used the **selections** feature of D3.js, an efficient declarative approach to operate on any sets of DOM nodes. All the mathematic calculation is made by the math API provided by D3.js, for example, min, max, mean values of an array of numbers. I used D3.js to implemented the bar charts in the List Analysis view all from scratch, for example, the animated transitions of the bars initialization, scaling, linked highlighting triggered by hovering over the bars and drawing all the SVG figures, including tooltips, axes, bars, average lines and so on.
- **Parcoords.js.** Parcoords.js is a d3-based parallel coordinates plotting library. It is used to implement the two parallel coordinates panels in both two views. All the interactive features I implemented for the parallel coordinates panels (for example, dragging, brushing, highlighting, colouring, progressive rendering, adding new axis for String Match attribute) are built on top of Parcoords.js API.
- **SlickGrid.js.** SlickGrid.js is one of the best libraries to implement fast JavaScript spreadsheet rendering [7]. It is used to implement all the two tables showing the detailed data of the keywords in the two views. The features I made for the tables (for example, hiding unwanted columns, row highlighting, row selection, table cell selection, content formatting, drawing the traffic percent bars, sortable columns) are built on top of SlickGrid.js API.
- **Chart.js.** Chart.js is a library to draw simple HTML5 charts using the <canvas> tag [2]. It is used to draw the search volume trends line charts in the compare mode.

Besides the API I used from the above libraries, I built everything in this visualization system from scratch by myself, including all the layout, functionalities of all the buttons, logics and control of the workflow. The tools I used and skills learning resources:

- **Sublime Text 2.** It is a sophisticated developing environment for coding HTML, JavaScript, and CSS. I wrote all the code in Sublime Text 2.
- **Git.** Git is used as the version control system of this project. It is free and open source. It handles perfectly with branching when I was working on different features in parallel. Based on the records on Git, this project has iterated to the newest 27th version.
- **Google Chrome Developer Tools.** The Chrome Developer Tools are a set of tools built into Google Chrome browser. Almost every front-end developer loves Google Chrome Developer Tool. I used this set of tools for all the debugging and inspecting the elements automatically generated by the code.
- **Spectrum.** Spectrum is an application on Mac OS X for creating and managing colour schemes. I used this Spectrum for colour scheme in compare mode.
- **Lynda.com.** Lynda.com provides online education for learning up-to-date technologies and skills. Since this is the first front-end project I have ever did, I took courses of JavaScript, CSS, HTML5, and d3.js on Lynda.com.
- **SEMrush.com.** All the keywords report data in this project is from SEMrush.com. The support specialist Mike Issac helped by answering some questions I asked about their data details.

6 Scenarios of Use

Terry has been running the domain hotels.com for more than a year. This website provides services for travellers, including search and reservation services for hotels, cars, and flights. Now he wants to start advertising his domain on Google search. He had a hard time choosing keywords when he set up the first ads campaign in Google AdWords, because very limited information is provided in the step of choosing keywords. He doesn't want this ads campaign become an expensive flop with little profit due to poorly chosen keywords. He knows that SEMrush.com provides much more detailed data for keywords. He wants to use KeywordVis visualization system based on data from SEMrush.com to effectively discover and analyze keywords in order to start his new ads campaign.

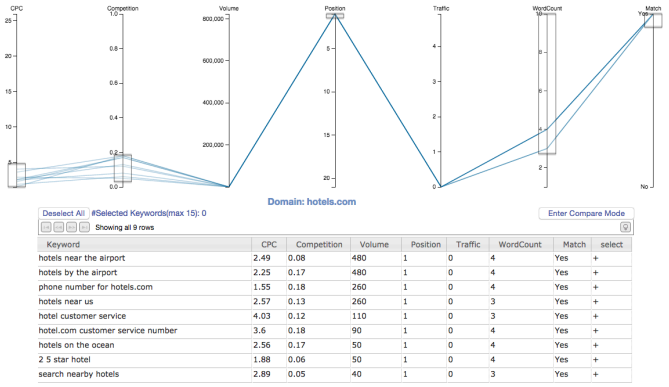


Fig. 17. Filtering keywords.

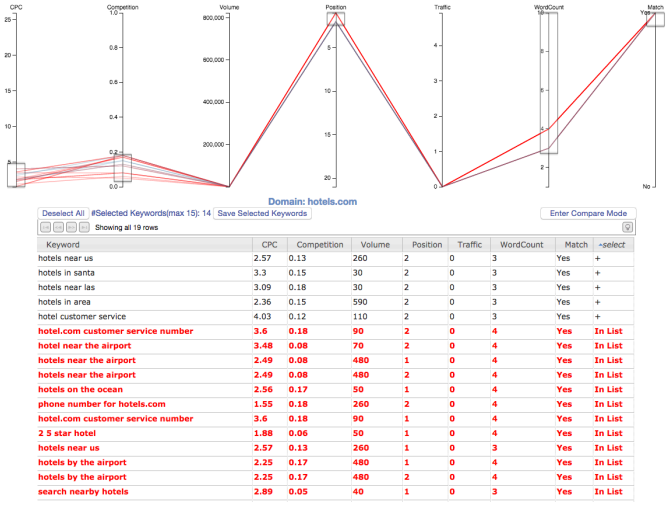


Fig. 18. Selecting keywords.

6.1 Scenario 1: Discover Long-tail Keywords

Opening the KeywordVis application, Terry lands on the List Discover view presenting all the 5,574 keywords in both a parallel coordinates

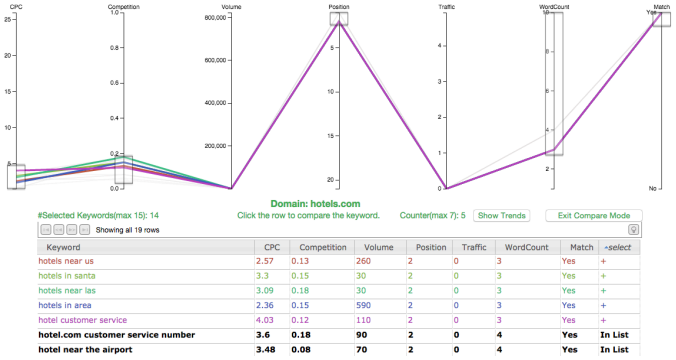


Fig. 19. Comparing keywords.

Fig. 20. Comparing keywords with search volume trends.

panel and a table. He wants to find the long-tail keywords containing the string "hotel". First, he adds an additional attribute to indicate whether "hotel" is contained by the keywords. Then he starts filtering the keywords by brushing and dragging on the axes. He explores the keywords with filters on CPC, Competition, Position, Word Count, and String Match as shown in figure 17. All the remaining 9 keywords look good for him so that he selects them all. Now he still want to find 6 more keywords. Then he changes the filter on Position attribute to include one more value. There are 10 more keywords. Terry sorts the Word Count column in the table and select additional 5 keywords with most Word Count values.

Then Terry wants to find the last keyword in the remaining 5. They are only different in CPC, Competition, and Volume attributes. He sorts the select column to put the 5 keywords as the top 5 rows shown in figure 18. He enters the compare mode and select all the 5 keywords by clicking the 5 rows. They are coloured differently. All the other keywords are temporarily coloured black so that Terry can focus on these compared keywords as shown in figure 19. By observing the 5 lines in parallel coordinates panel, He eliminates "hotel customer service" for its much higher CPC. He clicks "Show Trends" to explore how search volume trends for these five keywords changed in the past 12 months. Terry finds out that the trend for "hotels in santa" kept decreasing from 5 months ago shown in figure 20. So he eliminates this keyword as well. Among the remaining three, "hotels in area" and "hotels near us" both have strong growing trends. Meanwhile, "hotels in area" has lower CPC, same competition, and higher volume. So he finally selects "hotels in area" as the last keyword. After exiting the compare mode, Terry saves this list with name "hotels.com-list-1".

6.2 Scenario 2: List-to-list Analysis

Terry knows one of hotels.com's competitors is kayak.com. kayak.com has running ads campaign on Google search for a while. He wants to analyze the difference between his "hotels.com-list-1" list with the current list of keywords kayak.com is using. Terry goes to the List Analysis view and chooses kayak.com from the competitor drop-down menu. All the top 15 keywords with most traffic values are pre-selected. However, he finds out that the traffic value falls sharply to only 0.1 when it comes to the 15th keyword "vacation packages". Therefore he deselects this one and updates the competitor list to 14 keywords.

The bar charts and keyword lists are shown in figure 21. By observing the distribution and average lines shown by the bar charts, Terry finds out:

1. The keywords in his list contains much more words than the kayak's list.
2. The competition level and search volume of his list are much lower than the kayak's list.
3. The distribution and average value on CPC of the two lists are pretty similar.
4. All the values of traffic attribute are zero for the keywords in Terry's list.

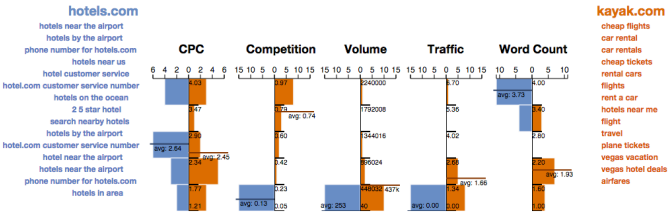


Fig. 21. Bar charts and keyword lists.

The value of traffic attribute is actually the ratio of the traffic brought by the keywords over the total traffic. Almost all the traffic is occupied by only few keywords so that the ratio of traffic brought by all the other keywords are almost zero. However, those few keywords are not long-tail keywords. This is the reason of point 4. Referring to the three features to evaluate the profitability of a keyword introduced in section 3.2.1, point 2 and 3 prove that this list of keywords has relatively low Search Volume and Competition, point 1, 4 and the fact they all have position 1 or 2 prove that the keywords in this list have high relevance to hotels.com. Therefore the keywords in Terry's list are long-tail keywords.

Terry's list of long-tail keywords allows the ads to reach people who matters most of his business. Moreover, his list is quite different from kayak's list in terms of Competition and Word Count, thus his website can avoid direct competition with kayak. This is almost a perfect keywords strategy for medium-small advertiser.

Terry wants to know the corresponding keywords represented by the top blue bar in the CPC vertical histogram, he hovers over that bar, all the corresponding keywords are highlighted in the list on the left side and in the parallel coordinates below.

7 Discussion

The final implementation of KeywordVis adequately supports the four tasks in section 3.2.2. Based on the data from SEMrush.com, it can help medium-small advertisers effectively find profitable long-tail keywords to set up new ads campaign on Google AdWords. In this way, they can avoid investing money in a lot of potentially profitable keywords and finding out only some of them can bring profitable clicks after a while. It also allows medium-small advertisers to make list-to-list analysis in order to make sure whether the keywords they select

are coherent with their advertising strategies, for example, not directly competing with other same-level or much bigger competitors and only reaching their target users. We hope KeywordVis can actually help the thousands of medium-small advertisers to grow their business.

7.1 Limitation and Future Work

There are still some limitation in KeywordVis currently:

1. For the current version of KeywordVis, users have to manually export the CSV files of the keyword report from SEMrush.com. This step should be simplified and embedded as a part of KeywordVis interface so that users can easily choose whatever domain within KeywordVis.

2. When brushing and dragging on the axes in the parallel coordinates panel, users could operate more accurately if the max and min values of the filters are shown. There is no existing API in the latest version of Parcoords.js library to support this feature. We would implemented it if we have more time.

3. The width of the entire layout of KeywordVis is fixed. So we recommend users to set the width of the browser to 1280 pixels. If the browser window is wider than 1280 pixels, the layout is aligned to left. This is because of a technical limitation. Using the Parcoords.js library, developers have to put the parallel coordinates panel in an absolute position within the browser window, otherwise the different components will fall apart.

4. All the keyword lists users have saved are stored within a single session supported by the HTML5 local storage technique for now. However, if users refresh or reopen the KeywordVis webpage, local storage is cleared. We could let users always access the lists they have ever saved by making a login system and a database on the server side if we have more time.

8 Conclusion

I built KeywordVis, a tool to visualize the ads keyword data from SEMrush.com in order to help medium-small advertisers to effectively find profitable long-tail keywords and analyze keyword lists to set up their ads campaign on Google AdWords. There are two views in KeywordVis: List Discover view for 1) deriving new attributes to indicate relevance, 2) determining a list of keywords, 3) comparing the attribute values across keywords, List Analysis view for 4) conducting list-to-list analysis with competitor's keywords. KeywordVid is already accessible online at `http://www.cs.ubc.ca/~yingsaid/VIS/KeywordVis/`. It would be interesting to keep improving KeywordVis so that it can be well used as a great supporting tool in the step of choosing keywords when advertisers set up ads campaign in Google AdWords.

Acknowledgments

The author wishs to thank Professor Tamara Munzner for her suggestion and feedback throughout the project. The author wants to thank Terry Guo for discussing JavaScript problems with him. The author also wants to appreciate Xinxin Zhang for his help on report refining.

References

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[2] N. Downie and other contributors. Chart.js. Online available at `http://www.chartjs.org/` [accessed 27-April-2015].

[3] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 127–130. IEEE, 2002.

[4] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. *STAR Proceedings of Eurographics*, 2013:95–116, 2013.

[5] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[6] A. Inselberg. Multidimensional detective. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 100–107. IEEE, 1997.

[7] M. Leibman and other contributors. Slick grid. Online available at `https://github.com/mleibman/SlickGrid` [accessed 27-April-2015].

[8] M. Redford. Understanding the fundamentals of keyword research. 2014. Online available at `http://www.keywordeye.com/blog/keyword-research-fundamentals/` [accessed 27-July-2014].

[9] R. Rosenbaum, J. Zhi, and B. Hamann. Progressive parallel coordinates. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 25–32. IEEE, 2012.

[10] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 105–112. IEEE, 2003.