# Faculty of Informatics and Design

## Cape Peninsula University of Technology (CPUT)

**This document is a working paper, summarizing a presentation at a conference. It was delivered at the SAICSIT 2012 Academic Conference in Irene, Pretoria, South Africa, at the Masters & Doctoral Symposium. The abstract has been reviewed and accepted, and it has been edited for basic grammar and accuracy. Kindly consider this document as a working paper, to be used for basic referencing. It could be useful for research in the field of Website Visibility, Digital Marketing and Search Engines.**

# Understanding choices of generic and branded keywords during query generation

Author: Jean-Marie SABWA

Student-Number: 210255951

Supervisor: Prof M. Weideman (CPUT)

Subject: Website Visibility, Digital Marketing and Search Engines

Harvard Reference:

Sabwa, JM. & Weideman, M. 2012. Understanding choices of generic and branded Keywords during query generation**.** *Working Paper*. Presented as an M&D Conference Paper at SAICSIT 2012, Pretoria, South Africa. 1 October 2012. http://cput.academia.edu/MeliusWeideman/Papers

## ABSTRACT

The aim of this study is to investigate the forces and mechanisms driving the search engine during query generation. The analysis further aims to improve understanding and design a conceptual model of intervention which will improve the search business and provide strategic indicators which can boost the performance of online businesses. Search engines are commercial entities that require revenue to survive. The most prevalent revenue stream for search engines is sponsored search, where content providers have search engines service their links to users in response to queries or in a contextual manner on relevant websites. In exchange for providing this service, content providers pay search engines based on the number of clicks generated by the relevant advertisements. Web sponsored search remains one of the most profitable business models in the world of search. It accounts for the main income of the highly successful Google, and generates revenues of at least 25 billion dollars per year.

A key technical challenge in sponsored search is to select advertisements that are relevant to the user's query. Identifying relevant ads is challenging since queries are usually very short, and because users, consciously or not, choose terms intended to lead to optimal Web search results and not to optimal advertisements. Furthermore, the ads themselves are short and usually formulated to capture the reader's attention rather than to facilitate query matching.

The objective of this research is to investigate and identify the search engines users' preference between generic and branded keywords during query generation. It has been established that less than half of search engine users with a university qualification select search engine advertising results as being more relevant than organic results. This research sets out to establish a division between branded and generic keywords as generated by search engine users.

The research methodology includes a survey of relevant literature and a questionnaire. The questionnaire was designed, and a pilot study was done to test its accuracy. The final version of the questionnaire was launched on a busy South African website. Statistical analysis was done, based on the categories; gender, age, educational qualification, race, Internet experience, working field and others.

The results provide a clearly different pattern of preference between generic and branded keywords. Further investigation is in progress to finalize these findings. In this paper, the preference of keywords (branded or generic) in sponsored search is outlined. Clear patterns were discovered, closely linked to education, age, and other factors. Conclusions include that online advertisers should consider these results, as they reflect the views of a proportion of search engine users, and they may be generalized. These results may be used by business modelers, content providers, and researchers. It is considered to be a unique form of information retrieval – not just a mode of advertising for web owners.

## 1. INTRODUCTION

The Internet is accessible worldwide, piping data and providing companies with a platform for their advertisements, online shops and services. It has become one of the fastest growing technologies in the world and the competition between the corporations is growing in almost the same manner. It is important  that  customers have to find website easily to satisfy their information need. Most customers use a search engine to find information or trade online, therefore website visibility become important, it used for business as well as for learning (Daniel 2008).

Search engines are commercial entities that require revenue to survive. The most prevalent revenue stream for search engines is sponsored search, where content providers have search engines service their links to users in response to queries or in a contextual manner on relevant websites. In exchange for providing this service, content providers pay search engines based on the number of clicks. This implies that online advertising budgets are expected to increase as well (Pfeiffer and Zinnbauer 2010). Recently, Web sponsored search remains one of the most profitable business models. It accounts for the survival of major search engines (Google and Yahoo!), and generates revenue of at least 25 billion dollars per year and rising (Graepel *et al* 2010).

Web search engines provide information access to millions of users per day. For many people, search engines are now the primary method for finding information, news, products and services, according to a report on Internet usage (Madden 2006). Given this importance, there is no attention being paid to search engine user's preferences of keyword retrieval.

Search engines offer two types of results on a search engine results page (SERP): sponsored and non-sponsored results (Kosin *et al 2007*). Finding the correct keywords in sponsored search is an increasingly important issue for providing a high return on investment. However, preference of search engine users and adversarial techniques have received little attention in the research community. This lack of consideration is surprising given that the positive feedback effect of user's preference on the sponsored search process may have greater implication in the business targeting categories.

A key technical challenge in sponsored search is to select ads that are relevant to the user's query. Identifying relevant ads is challenging since queries are usually very short, and because users, consciously or not, choose terms intended to lead to optimal Web search results and not to optimal ads. Furthermore, the ads themselves are short and usually formulated to capture the reader's attention rather than to facilitate query matching.

The focus of this research is to investigate and retrieve the variables which influences search engines users' preference between generic and branded keywords during query generation. The forces and mechanisms driving the search engines need to be understood by online busineses. This analysis is to improve the understanding, and may lead to a conceptual model of intervention which will improve the search business and lead up some strategic indicators that can boost the performance of online businesses. This research sets out to establish a trend in the direction of variables that have a clear division between branded and generic keywords by search engine users.

## 2. OTHER RESEARCH

### 2.1 Keywords

The use of descriptive keywords in a website for search engines to extract information is important. However, choosing the right keywords, and knowing where to locate them is not that simple. Weideman and Kritzinger (2007) indicated that the best locations to place keywords in the body text of a website are towards the beginning. Chambers (2005) is more specific by showing that the position of the keywords at the beginning also has the practical side effect - most users do not read a complete site, but rather scan for the words they are looking for. So if a user or customer find the words he/she searched for quickly, the chances that they will remain on the site are higher.

Ghose and Yang (2009) modelled the relationship between different sponsored search metrics such as click-through rates, conversion rates, cost per click, and ranking of advertisements. They quantified the relationship between various keyword characteristics, position of the advertisement, and the landing page quality score on consumer search and purchase behaviour as well as on advertiser's cost per click and the search engine's ranking decision. Among other important discoveries, they found that: “keywords that have more prominent positions on the search engine results page, and thus experience higher click-through or conversion rates, are not necessarily the most profitable ones—profits are often higher at the middle positions than at the top or the bottom ones". This last finding encourages this research to investigate the factors influencing the user's choices.

Yao and Mela (2011) proposed a dynamic structural model as a foundation to explore how the interaction of various agents (searchers, advertisers, and the search engine) in keyword markets affects consumer welfare and profits. They included in the model variables like: consumer search and clicking behaviour, advertiser bidding behaviour, and search engine information such as keyword pricing and website design. They found evidence of dynamic bidding behaviour. Advertiser value for clicks on their links averages about 26 cents. Only about 10% of consumers produce 90% of the clicks. They also found that a switch from a first- to second-price auction results in truth telling (advertiser bids rise to advertiser valuations). However, they also state that a second-price auction has little impact on search engine profits and these tools, by reducing advertising exposures, lower advertiser profits by 2.1%.

Additionally these authors state that the repeated use of a keyword will rank a webpage higher than if the keyword appears only in one location. But Chambers (2005) claims that using a keyword more than twice could be viewed as a potential spamming technique and the crawler could blacklist that webpage. Chambers also contend that the best way for a webmaster to choose keywords is to visualize how a user would create a search query and then to ensure that this search query is on the webpage. From this point of view, the current analysis is of great help, because webmasters and website owners should have the best understanding of user's choices. In these investigations it needs to be determined what influences the choice of keywords being branded or generic. All the above studies were done on the choices already made by users.

### 2.2 Filename and Directory Naming Conventions

As noted in Section 2.1, the keywords are of great importance in every aspect of online business; they need to be understood in structures, positions as well as in a filename. Weideman (2007) showed that keywords should be used in all HTML page names except the homepage (index.html), and the filename should not be longer than 30 characters. For example, a sensible filename could be "learn-english.html", if the webpage is about learning English. So the user and the search engine will find two keywords in the filename of the webpage.  An underscore should not be used for separating the keywords, since some search engines will read it as only one keyword; for example (“learn_english.html”) might be considered as (“learnenglish”) (Anonymous 2008:1).

An additional naming convention to consider is when directories are named. If a website sells car-parts in different categories such as tyres, wheels and windshields, these different categories should be placed into different directories. For example:
- http://www.carparts.com/windshields/gmc-windshields.html,
- http://www.carparts.com/tyres/ford-tires.html)
(Anonymous, 2008:1).

Thus the file and the directory names will form part of the URL and the search engine will consider this fact as part of their algorithm.

## 3. METHODOLOGY AND RESULTS

The research methodology includes a survey of relevant literature and a questionnaire done with a sample of some search engine users. Due to the fact that South Africa's leading search engine is Ananzi.com, the authors decided to focus on Ananzi.com. Statistical analysis, based on the categories; gender, age, educational qualification, race, Internet experience, working field and others was used to identify different trends of choices between branded or generic keywords.  The author decided to execute  this experiment to determine whether or not users of webpage have preferences on keywords based on their background.

### 3 .1. Analysis based on questionnaire survey

#### 3.1.1 Data selection

The following dependent variables were considered as primary measures that influence keyword choices:

##### 3.1.1.1. Gender

We considered gender in our analysis to investigate if it has a significant influence on keyword choices.

##### 3.1.1.2. Age

The variable age also was considered to be a predictor in the choices of key words. The data were classified as follow:

- Under 20 years
- Between 20 and 30 years
- Above 30years old

#### 3.1.1.3. Highest qualification achieved

It has been established that only 47.1% of search engine users with a university qualification select search engine advertising results (PPC) as more relevant (Neethling 2007). So, one can conclude that university qualification has an impact on how users generate queries. This variable was considered also to predict the choice of key words, considered as follow:

- Lower than matric (grade 12)
- Matric
- Bachelor degree
- Master's degree
- Doctorate degree

### 3.1.2. Keywords

The researchers selected the option which would be used as a search query to find information on perfume, computers, cameras, loans, cellular handsets, ovens, music CD/DVDs, hotel accommodation in Cape Town, second hand cars and companies offering home or car insurance. The choice was to be made between generic and branded key words during a query on a search engine on the above categories.

## 3.2 General CHAID Model

Two types of keywords are being considered. The authors wanted to investigate the impact of the above-mentioned variables on the choices of keywords. The CHAID tree algorithm can classify the data so that from these predictors it will be possible to determine the type of keyword that is likely to be classified. After eliminating empty cells, 158 rows remained, against 14 columns of predictors (of above-mentioned variables) which were used for the analysis.

Initially all the variables were included in the CHAID model to see how well the model works. **Figure1.1** indicates how the factor "age" had an impact on the choice of keywords. An example of loan branded and generic key words was tested.

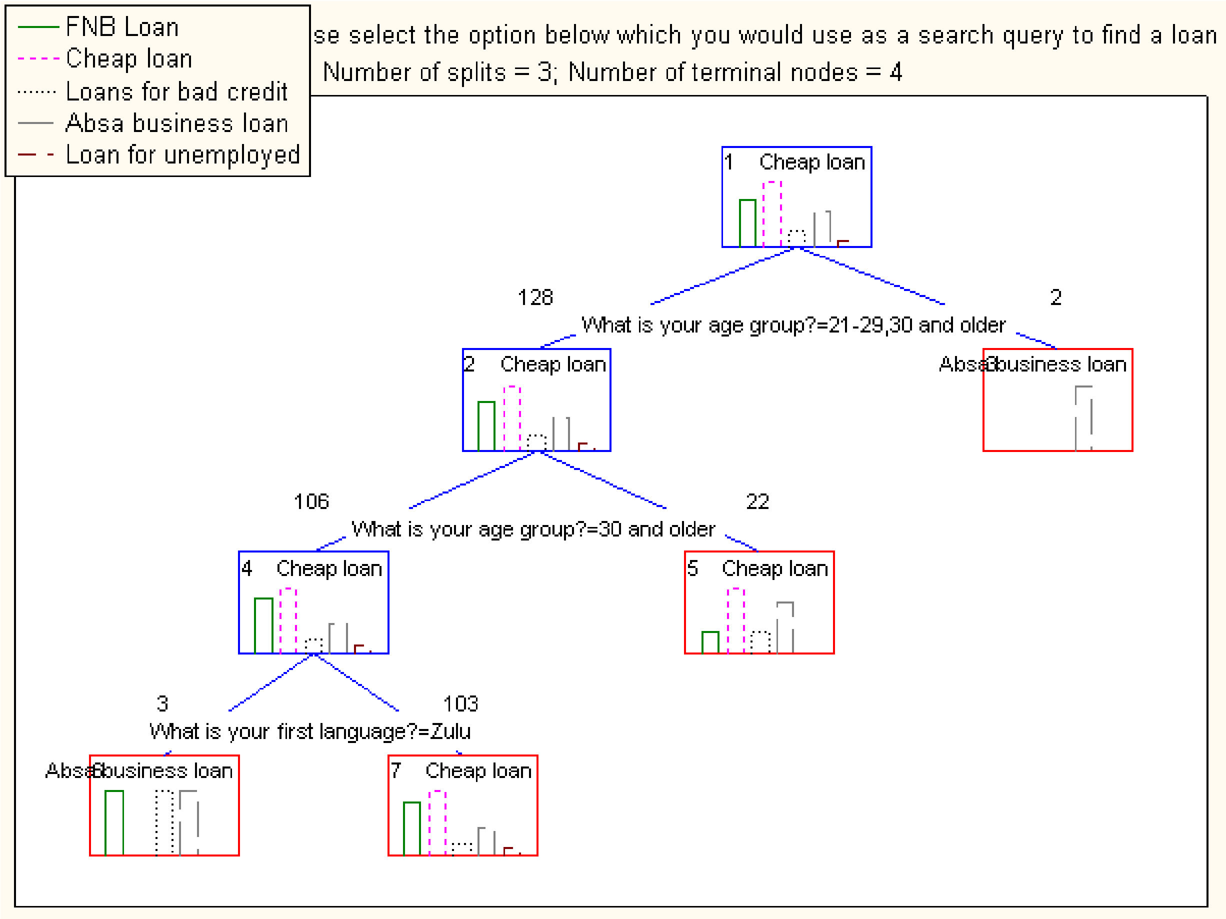
FNB Loan
Cheap loan
Loans for bad credit
Absa business loan
Loan for unemployed
se select the option below which you would use as a search query to find a loan
Number of splits = 3; Number of terminal nodes = 4
1 Cheap loan
128
2
What is your age group?=21-29,30 and older
2 Cheap loan
Absa business loan
106
22
What is your age group?=30 and older
4 Cheap loan
5 Cheap loan
3
103
What is your first language?=Zulu
Absa business loan
7 Cheap loan

**Fig 1.1**

The trend is clear - people more advanced in age (above 30 years) have a preference for generic as opposed to branded keywords, while Sabwa and Weideman (2010) found that 81% of young consumers prefer natural above sponsored results. In this research, an equal number of choices between branded and generic were offered. The first split shows the choices according to the type of keywords as follows:

- The generic keyword of "cheap loan" was associated with the age factor.
- The others factor, like language, had little impact.

The second split shows clearly that:

- In the group of people who chose "cheap loan" as a keyword, the majority were above 30 years old (106 out of 128).

- The smaller group was formed by people under 30 years old.

The third split reveals that the language had little impact on the choices.

In order to investigate further the impact of age on the choice of keywords, a series of variable graphs were constructed. The variable "age" combined with gender and mixture of different keywords was plotted. **Figure 1.2** illustrates the outcomes.

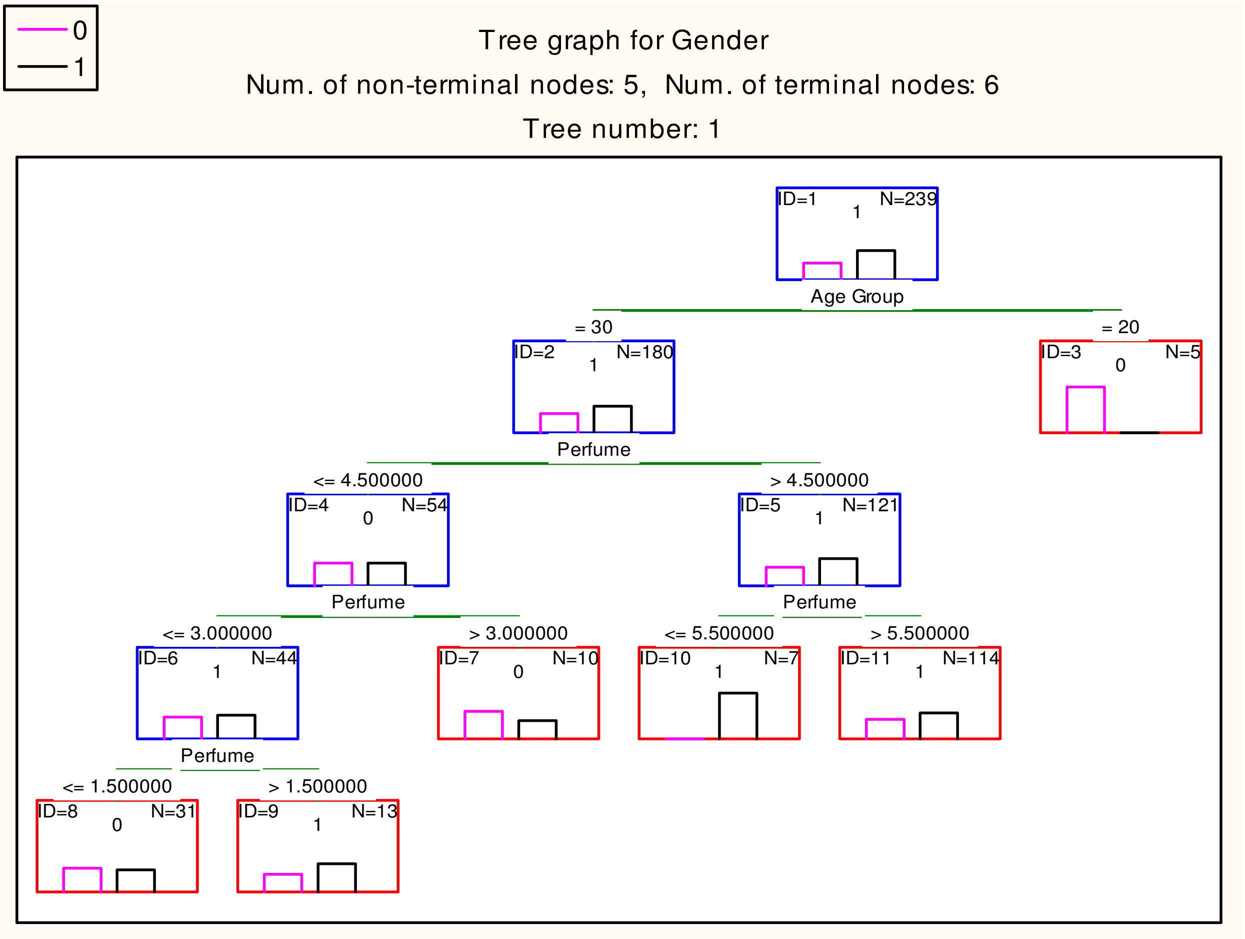

**Figure 1.2**

It is clear from **Figure 1.2** that the variable "age" has an impact on the choices of keyword when associated with gender. The split indicates the difference in choices according to age groups.

- The group of people chose a generic keyword “perfume“ as a preferred choice when searching online for fragrances.
- Most of the respondents on the keyword "perfume" were above 30 years old and male dominated.

# 4. CONCLUSION

This analysis was used to investigate the impact of these variables (gender, search engine experience, level of education, age, etc) on the choices of different types of keywords. It was also used to enable identification of significant differences in the behaviour patterns of different types of targeted potential online customers. The factor age and level of education have an impact on how these choices are made. From the CHAID results, the following were identified:

- There is a clear pattern on how users choose their keywords to find information online
- The choice of keywords is influenced by the factor age and level of studies.

When assessing the statistical results on an individually basis it was determined that age, gender and level of education has the largest impact on the user choices for search engine query generation.

Almost eight categories of items were selected randomly for the keyword investigations, and it was found that age was a stronger influencing variable than gender. However, the survey needs to be expanded to other popular exposing platforms to comfirm or refute these findings.

## 5. REFERENCES

Anonymous. 2008. *File Naming Conventions - SEO Rules For Naming Web Pages - How To Make A Website*. http://www.makeawebsitespot.com/file-naming-conventions.htm [09 November 2012].

Chambers, R. 2005. Search engine strategies: A model to improve website visibility for SMME websites. Unpublished Master's Thesis (MTech, IT), Cape Peninsula University of Technology, Cape Town.

Daniel, W. 2008.The Importance of Internet in Education. http://www.associatedcontent.com/article/ [19 November 2012].

Graepel, T., Candela, J.Q., Borchert, T.and Herbrich, R. 2010. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. *Proceedings of the 27th International Conference on Machine Learning,* Haifa, Israel: 13-20.

Ghose, A. and Yang, S. 2009. An empirical analysis of search engine advertising: sponsored search in electronic markets. *Management Science,* 55: 1605-1622.

Kosin, I. Yasushi, K. and Victor. T.J. 2007. Generic and Branded Advertising in Markets with Product differentiation. *Journal of Agricultural & Food Industrial Organization,* 5(1).

Madden, M. 2006. *Internet Penetration and Impact*. http://www.pewinternet.org/PPF/r/182/report_display.asp. [9 October 2012].

Neethling, R. 2007. Search engine optimization or paid placement systems-user preference. Unpublished Masters Thesis (Mech, IT), Cape Peninsula University of Technology, Cape Town.

Pfeiffer, M. and Zinnbauer, M. 2010 Can old media enhance new media? How traditional advertising pays off for an online Social Network. *Journal of Advertising Research,* 50(1).http://www.webanalyticsassociation.org [19 November 2012].

Sabwa, J.M. and Weideman, M. 2010. Paid search versus organic results: young consumer preferences. *Proceedings of the 12 th World Wide Web Applications*, Durban, 21-23 September. www.zaw3.co.za [14 November 2012].

Weideman, M & Kritzinger, W. 2007. Key word placing in Web page body text to increase visibility to search engines. *South African Journal of Information Managment, 9(1).* www.sajim.co.za [23 October 2012].

Weideman, M. 2007. Best Practice Summary – Search Engine Optimisation updated 06. November 2007.

Yao, S and Mela, C.F. 2011. A Dynamic Model of Sponsored Search Advertising. *sInforms online-Marketing Science*, 30: 447-468.