Search Engine Metrics to Discover Terms Characteristic of a Database of Images with Captions

Michael Oakes, Lynne Hall
Department of Computing, Engineering and Technology
University of Sunderland
Sunderland, SR6 0DD, United Kingdom
{michael.oakes,lynne.hall}@sunderland.ac.uk

Abstract

In this paper we make use of search engine metrics to discover the terms, which are most typical of a web-based image collection, as opposed to those terms used by search engines in general. Knowing which terms are typical of a cultural data collection is of interest in itself, and these terms can also be used to draw additional visitors to these websites through search engine optimization (SEO). To find terms suitable for paid placements (Adwords), the metric we propose is relative cost per click (RCPC), the ratio between Google's Cost-per-Click (CPC) and the frequency of the term in the captions of the images in the collection. To find terms suitable for metadata to improve the matches between user queries and the image website, we propose a variant of the term frequency-inverse document frequency (tf.idf) metric commonly used in information retrieval.

1 Introduction

It is important for a web-site to appear on the first page of a search engine "hit" list, which typically consists of the 10 sites which best match the query, as most users look no further. Originally, search engine algorithms, based on the degree of match between the terms of the user's query and the words in the index to each web-site, were the sole determinants of rankings. Nowadays, however, one can also get a link to one's web-site on the first page by paid placement (also called CPC), whereby for a fee proportional to the number of users who click on that link, any query containing a specific bought term will cause your site link to appear in a special box on the first page. CPC fees are tied to the price of the relevant keywords (called "Adwords"), as determined by online auction (Xing and Lin, 2006). The advantage of the "algorithmic" approach is that searchers prefer these links as they are felt to be more objective than paid placements. However, using paid placements is now very popular, being cheaper than other methods of SEO or trying to move a web-site up the search engine rankings (Wiedeman, 2004). Advantages of paid placements include the targeted audience and the low cost per advertisement. The key to successful paid placement for the advertiser is to select cost-effective keywords. Keyword selection is also important when these terms are used as metadata tags to improve the web-site rankings in algorithmic searches. In this paper, we describe metrics for the automatic selection of keywords for both purposes, using the Aframe image collection[1] which contains both still and video images, each with an associated caption written by professional taggers. We distinguish four types of keywords: Adwords, image caption tags, metadata to be placed in the HTML of the homepage which links to the images themselves, and query terms input to the Aframe web site's internal search engine.

2 Metric for the selection of cost-effective terms for paid placement

We propose a metric for the relative cost of a single keyword for paid placement (AdWord), which is the ratio of Google's own estimate of

[1] URL: http://aframe.com

Short papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 131–134
Paris, 8–10 November 2011

the cost per click of that keyword, found using the AdWord tool[2] to the frequency with which that keyword has been used as a caption tag in Aframe's own image collection. Relatively cheap keywords will tend to have low Google CPC, but be used relatively frequently by the Aframe caption taggers – hence not greatly in demand by advertisers in general, but highly reflective of Aframe's own content. Frequency data for the 100 most-used caption tags was provided by Aframe. No particular semantic pattern was seen among the most cost-effective Adword keywords found by this method, which are listed in Table 1:

Term	CPC	Frequency	RCPC(x10^3)
Shot	1.97	11303	0.174
Right	0.66	3377	0.195
Very	0.22	1115	0.197
Up	0.87	4058	0.214
Camera	2.45	8869	0.276
No	1.16	3339	0.347
Laughing	0.51	1451	0.351
Monkey	0.77	1846	0.417
One	0.66	1334	0.494
Medium	2.32	4286	0.541
Unknown	1.53	2812	0.544
Out	1.77	3195	0.554
Sky	0.50	861	0.581
Daytime	1.66	2719	0.611
Pan	2.35	3711	0.633
Clapping	0.73	955	0.764
Female	1.38	1804	0.765
Interior	3.56	4616	0.771
Chatter	0.91	1134	0.802
Sound	2.96	3480	0.850

Table 1. The 20 most cost-effective terms for paid placement for Aframe.

According to this list, shot would be the most cost-effective AdWord. We will see in the following section that this word is also most typical of Aframe tags as opposed to Google index terms in general, and thus would be both an excellent AdWord and meta-tag term. Some of the list of 100 most popular Aframe tags were not available as AdWords: hand, smiling, standing, from, landing, crowd, trick, performing, adjustment, talking, left and close-up, as indicated by a CPC value of 0. Google's estimated CPC, which varies continually, was recorded on Monday 4/7/11 at 3pm.

Single words are not considered cost-effective AdWord keywords (Enge et al., 2010:171), but the method described in this section is also applicable to finding pairs or longer sequences of words. Abhishek and Hosanagar (2007) write that such "bought" keywords with very high volumes are expensive. However, it may be possible to find a larger number of keywords, all semantically related to the original "obvious" Adword, and generally consisting of more than one word. These would "generate the same amount of traffic cumulatively but are much cheaper". Approaches to finding many such "long tail" terms, include using a thesaurus such as WordNet[3]; finding terms which tend to occur in the same image captions as words already found, or using the WordTracker[4] tool, which can generate variants of a single "seed" phrase. In response to shot, WordTracker generated self shot, self shot mirror, police mug shots, criminal mug shots online, local mug shots, inmate mug shots, mug shots of people in jail, money shot, Britney Spears crouch shots, and I shot myself, but none of these are relevant to the Aframe collection.

3 Metric for the selection of meta-tag keywords

In this section we propose a metric to identify the best metadata keywords to be placed in the HTML of the Aframe homepage, to make the website appear higher up on the Google hit list for an algorithmic search. (In contrast, the metric described in Section 2 found candidate keywords for paid placement). These optimal metadata keywords were found using a version of the term frequency-inverse document frequency (tf.idf measure commonly used in information retrieval for weighting (estimating the relative importance of) index terms with respect to a web-site by a search engine. One version of the formula for the tf.idf measure (Meadow et. Al., 2000: 217) is:

$$w_{ij} = tf_{ij} . \log_2 (N / D_i)$$

where w_{ij} is the importance of word i in web-page j, tf_{ij} is the number of times word i is found on web-page j, N is the total number of web-pages indexed by the search engine, and D_i is the number of those web-pages which contain term i.

[2] URL: adwords.google.co.uk/select/KeywordToolExternal

[3] URL: wordnet.princeton.edu/wordnet/download

[4] URL: wordtracker.com

The idea is that a word is important to a web-page if it not only occurs frequently on that web-page, but occurs in few other web-pages. The problem of finding optimal keywords for the Aframe web-site is slightly different. We want to find those keywords which are frequently used by users of the Aframe search engine, but which are relatively infrequently used by users of Google (a “typical” search engine). To find the words most used by the users of the Aframe search engine, we would ideally look at the data in Aframe’s search logs (transcripts of interactions between previous users and the search engine). Since this data was not immediately available, we used a surrogate measure – the frequency with which each caption tag had been used by the Aframe taggers in indexing the image collection for the Aframe search engine. Clearly the most often tagged concepts were important to the Aframe search engine and its users, and so we used the frequencies of the 100 most popular tags as the tf values in the formula. However, we also wanted to filter out those keywords, which were commonly submitted to search engines in general. To estimate this, we estimated D for each keyword by submitting it to Google, and recorded the number of hits (the number of web-sites indexed by that keyword) that subsequently appeared just below the query bar. N was the estimated total number of web-sites indexed by Google, taken to be the number of web-sites indexed by the (25.27 billion). The set of keywords most valuable to Aframe as opposed to Google in general are given in Table 2. This data shows that the most Aframe-specific keyword is ‘shot’, and a number of other high scoring words also pertain to types of camera shot. One implication of this is that a strength of the Aframe image collection is that the image captions contain details not only of what the picture is of, but the kind of camera shot with which it was taken. This could be exploited if the Aframe web site were to be meta-tagged with such terms as shot, camera, exterior, interior, pan or close-up which would draw in users who mention types of camera shots in their general search engine queries. Less importantly there are three highly-scoring keywords related to monkeys – suggesting that people interested in these animals could also be directed to the Aframe site by the inclusion of monkey, monkeys and chimpanzee in the meta-tags. Words like this are useless (and have a weight of 0), because they appear in every Google website. Since the number of hits per query term (postings) data used in this experiment varies over time, in a larger study one should sample the data at different time intervals, and take the mean of the inter-quartile range as the “average”.

Term	**Tf$_{ij}$**	**D$_i$ (x10^{-3})**	**w$_{ij}$**
Shot	11303	1060	51714
Camera	8869	2040	33201
Exterior	4353	635	23134
Daytime	2719	89	22158
Interior	4616	1160	20519
Pan	3711	664	19483
Close-up	3403	1020	15758
Chimpanzee	1356	9	15533
Medium	4286	2070	15471
Zoom	3205	1060	14663
Unknown	2812	1050	12904
Sound	3480	2050	12610
Talking	2526	939	11998
Walking	2301	713	11844
Monkey	1846	356	11351
Chatter	1134	51.5	10136
Wide	2623	1820	9955
Adjustment	1375	197	9629
Interview	1904	833	9373
Motion	1770	660	9308

Table 2: The 20 most recommended terms for meta-tags by the tf.idf measure

4 Other metrics for finding terms useful for search engine optimization.

As mentioned in section 3, the frequencies with which keywords have been used in the image captions tell us how important the keywords are to the Aframe collection. The Aframe web site has its own internal search engine, and in future we will maintain a search log: a list of all queries entered into the engine (Garcia, 2007). We will then be able to find the frequencies of query terms in this search log to better estimate which keywords (search terms) are most important to the users of the Aframe internal search engine.

Various authors have proposed metrics for evaluating web-sites or advertising campaigns. These can be used to rank keywords by the effect that a single keyword might have on the number of visitors, if as either an Adword or a meta-tag. These metrics include click through rate (the proportion of search engine users who see a web link who click on it) (Joachims, 2002), conversion rate (proportion of people visiting a web-site who actually download what is on offer, such as an Aframe image) and "bounce" rate (proportion of people who leave a web-site after 5 to 60 seconds without performing any other action – this is inversely related to user satisfaction). All these can be derived from search logs (Sculley et al., 2009).

In an evaluation of automatically-suggested keywords for SEO by Joshi and Motwani (2006), each suggested term was given two ratings: relevance (as determined by human judges) and non-obviousness (not containing a seed keyword or its grammatical variants as found automatically using a Porter stemmer). To compare keyword generating techniques, we can then use average precision, average recall (using the union of all relevant keywords from a the range of techniques being compared) and average non-obviousness (Joshi and Motwani, 2006).

5 Conclusion

We have described two metrics for term extraction that use search engine technology. One advantage of these techniques is that they are largely language independent. The only "English-only" tool that we have made use of is the WordTracker tool which finds less common phrases semantically-related to an original single "seed" term. Statistical measures based on the contingency table, such as the chi-squared test, have been used for determining the vocabulary characteristic of one body of text as opposed to another. The metrics described here use a related principle, since they aim to extract the terms typical of one collection of images (such as the Aframe collection), as opposed to those used by search engines in general (as exemplified by Google). However, the resulting weights are not so much related to statistical significance, but more directly related to a cost-benefit analysis of their use in search engine optimization.

Acknowledgments

This work was partially supported by European Community (EC) and is currently funded by the ECUTE project (ICT-5-4.2-257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

Vibhanshu Abhishek and Kartik Hosanagar. 2007. Keyword generation for search engine advertising using semantic similarity between terms. Proceedings of the 9th International Conference on Electronic Commerce (ICEC '07): 89-94.

Eric Enge, Stephan Spencer, Rand Fishkin, Jessie C. Stricchiola. 2010. The Art of SEO – Mastering Search Engine Optimization. O'Reilly, Beijing, China.

Steven Garcia. 2007. Search Engine Optimisation using Past Queries. Ph.D. Thesis (Abstract), RMIT University, Melbourne, Australia.

Thorsten Joachims, 2002. Optimising search engines using clickthrough data. Proceedings of the Special Interest Group in Knowledge Discovery in Databases (SIGKDD 02), Edmonton, Alberta, Canada: 133-141

Amruta Joshi and Rajeev Motwani. 2006. Keyword Generation for Search Engine Advertising., Proceedings of the 6th IEEE Conference on Data Mining – Workshops (ICDMW '06): 490-496.

Charles T. Meadow, Bert R. Boyce, Donald H. Kraft. 2000. Text Information Retrieval Systems. Academic Press, San Diego. Second Edition: 217.

D. Sculley, Robert Malkin, Sugato Basu, Roberto J. Bayardo. 2009. Predicting bounce rates in sponsored search advertisements. Proceedings of Knowledge Discovery in Databases (KDD 09), June 28 – July 1, Paris, France: 1325-1334.

M. Weideman. 2004. Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results. Informatics and Design Papers and Reports, Paper 77. http://dk.cput.ac.za/inf_papers/77

Bo Xing and Zhangxi Lin. 2006. The impact of search engine optimization on the online advertising market. 8th International Conference on E-Commerce (ICEC 2006), Frederickton, New Brunswick, Canada: 13-16.