

Technical Disclosure Commons

Defensive Publications Series

November 30, 2016

SYSTEM AND METHOD FOR AUTOMATIC NEAR EXACT KEYWORD CANONICALIZATION

Justin Lewis

Gavin James

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation

Lewis, Justin and James, Gavin, "SYSTEM AND METHOD FOR AUTOMATIC NEAR EXACT KEYWORD CANONICALIZATION", Technical Disclosure Commons, (November 30, 2016)
http://www.tdcommons.org/dpubs_series/335



This work is licensed under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/).

This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

SYSTEM AND METHOD FOR AUTOMATIC NEAR EXACT KEYWORD CANONICALIZATION

ABSTRACT: This disclosure relates to systems and methods for automatic near-exact keyword canonicalization. In particular, a keyword management system is provided that, based on keywords input by an advertiser, algorithmically derives a set of canonicalized IDs that represent the input keywords and near-exact formulations. The keyword management system also dynamically updates the collection of canonicalized IDs, as necessary, in response to new keywords input by the advertiser. The system is also configured to provide user feedback so as to educate the advertiser about the canonicalization of the entered keywords.

Advertisement auction systems allow for various types of matches for search terms (keywords) that advertisers can bid on to display ads including: exact match, broad match, phrase match, negative match, and broad match modifier. In an exact match, an advertisement is associated with keywords and the advertisement is displayed to a searcher when the searcher performs a search query that exactly includes the keywords associated with the advertisement. Advertisement auctions also support a broad match of keywords in which an advertisement is matched with a keyword and with variations of the keyword (“variation keywords”) in any order.

Exact match can be challenging for a user to use effectively. There are also challenges associated with effectively managing and implementing targeted advertising systems using broad-match allocation criteria as well. In order to address the challenges associated with managing strict exact match allocation criteria, “near-exact match” selection criteria generally considers spelling errors and plural forms of a keyword to be the same as the keyword (i.e. a match). In addition, near exact match can also consider different word orders, possessives, or even closely related synonyms to be the same keyword. Although “keyword” is used in the

singular form, a keyword can consist of multiple words individually referred to as “tokens.” For example, “women’s hats” is a single keyword that consists of two tokens.

As further described herein, the process of defining a common keyword for related, near-exact match keywords is referred to as “canonicalization.” Continuing the practical example, according to the disclosed embodiments, the keyword “women’s hats” can be canonicalized to a more generic representation, such as, “woman hat,” or another such representation. Preferably, the keyword input by the advertiser and all appropriate variations thereof (e.g., near-exact matches) are collected into the canonicalized representation.

The systems and methods described herein provide user feedback that educates the advertiser as to the canonicalization of the particular keyword or keywords entered by the advertiser. Advertisers currently enter words approximating the keywords that they intend to target and the keyword management and bidding software program identifies the near exact keywords and implements the near exact keyword selection process as a background process that is not visible to the advertiser. As a result, advertisers lack information sufficient to form an understanding as to what keywords and near exact keywords are actually used in connection with their original keyword input. While an alternative option is to provide a translation tool that displays the canonicalized versions of the advertiser’s input, this can often result in a choppy, broken-up workflow that can detract from the user experience and utility of the system.

The disclosed systems and methods provide a mechanism for automatically canonicalizing keywords input by an advertiser and managing keywords for use in keyword-based match auctions. In particular, the disclosed embodiments can be implemented to more effectively aid advertising keyword selection for an advertiser so as to increase the utility for the advertiser in a keyword match auction. In the application of the disclosed embodiments,

advertisers can type or add one or more keywords into a user interface element. A keyword management system is also provided that is configured to process the input keywords and algorithmically derive a collection of canonicalized IDs representing respective keywords. The keyword management system is also configured to dynamically update the collection of canonicalized IDs such that, each time the set of keywords the advertiser typed (or otherwise input) is adjusted, the collection of canonicalized IDs can be recalculated and updated.

FIG. 1 is a conceptual diagram of an environment 100 in which a keyword management system 110 is used to manage keywords associated with keyword-based match auctions. The environment 100 includes network 102, e.g., a local area network (LAN), a wide area network (WAN), the Internet, or a combination of them, that connects advertisers 104, user devices 106, search and auction engine 108 and the keyword management system 110.

Each advertiser includes a person, collection of people, or other entity who wishes to make information available to an audience. An advertisement (“ad”) is a vehicle through which the advertiser brings the information to the audience. Advertisers 104 submit keywords to keyword management system 110 through network 102. Queries 112 include user searches that include particular keywords 105 that are performed over network 102. A keyword includes words selected by an advertiser as relevant to an advertisement of the advertiser. User device 106 includes an electronic device that is under control of a user (e.g., a searcher) and is capable of sending queries 112 (e.g., search queries) over network 102. Example user devices 106 include personal computers, mobile communication device, and other devices that can send and receive data over the network 102. It should be appreciated that advertisers 104 similarly interact with the system 100 using respective computing devices.

Search and auction engine 108 matches queries 112 with bids for keywords and assigns advertisement slots to advertisers 104 by determining an advertiser that has a winning bid for keywords included in queries 112. Keyword management system 110 or auction engine 108 can also be configured to generate bids for one or more keywords. Keyword management system 110 can be any of a variety of computing devices capable of receiving/transmitting information to other computing devices and implementing the exemplary methods described herein.

Advertisers 104 submit targeting criteria including keywords 105 to the keyword management system 110 using a configuration user interface for configuring and managing respective accounts. The targeting keywords 105 can be stored in a keyword index 118, in association with an advertiser 104a or a particular content group or a content campaign.

Keyword management system 110 can include a keyword control module 130 configured to provide a graphical user interface (GUI) that can be displayed on a computing device of the advertiser 104a, for example, as a web page for managing keywords and campaigns. For example, the control module can provide an interactive GUI having a “keywords area” through which the advertiser 104a can add one or more keywords to a content group or campaign and manage the keywords, related bidding parameters and the like. The keywords area can also output relevant information relating to the keywords input by the advertiser. Keyword management system 110 can also include a keyword evaluator module 132 that is configured to evaluate and convert the input keyword(s) into a canonicalized keyword format.

A user feedback module 134 can also provide information relating to the canonicalization of keywords back to the advertiser 104a. For example, the feedback module 134 and/or the control module 130 can output a set of unique canonicalized IDs generated using the keyword

evaluator module 132. Accordingly, the canonicalized IDs can be presented to the advertiser in a GUI displayed by the advertiser 104a's computing device.

FIG. 2 is a flowchart of an example process 200 for managing keywords. The process 200 can be performed, for example, by the keyword management system 110. The process begins as one or more keywords for an advertising campaign are identified (202) by the keyword management system 110. For instance, the keywords 105 input by advertiser 104a are received by the keyword management system 110 over the network.

The one or more initial keywords are then transformed into a corresponding one or more canonicalized keywords (204). In particular, the keyword evaluator module 132, can match each initial keyword to an index of canonicalized keywords, which are also referred to as "canonical IDs." The canonical IDs and respective sets of one or more corresponding keywords can be maintained in a canonical IDs index 116. Preferably the canonical IDs maintained in the index 116 are defined and associated with respective sets of keywords such that all keywords that represent the same entity (e.g., misspellings of the same word, synonyms and the like) are represented by a single canonical ID.

Language processing algorithms can be used to analyze the keyword and, if necessary, transform the keyword into a normalized form that is more suitable for identifying the corresponding canonical ID. For instance, the keyword can be transformed using spelling or grammar considerations (e.g., placed into a singular and non-possessive format representing the entity). Keywords can also be canonicalized based on the context in which the keyword is used, such as a function of geography and language.

Because each canonical ID inherently represents several variations of a keyword such as misspellings, synonyms, re-orderings, and other near-exact formulations, after a single canonical

keyword is identified for each of the keywords input by the advertiser 104a, any repeated canonical IDs can be discarded (206) or otherwise ignored by the keyword management system 110 when implementing the campaign.

A salient component of the disclosed embodiments is that the set of unique canonicalized keywords (i.e., canonical IDs) can be automatically shown to the advertiser (208). Today, when an advertiser manages keyword-based targeting, the advertiser generally inputs a slew of different keywords (often these include misspellings and synonyms) and, when the advertiser inspects an ongoing campaign, are provided with a list of the set of keywords that were originally entered as targeting criteria. Currently, any use of canonicalization is implemented as an additive process, wherein the set of keywords that the advertiser provides are used for targeting and, in the background, the set of keywords is expanded by canonicalizing the keywords, and then broad matching is performed against the canonical versions. The disclosed embodiments, however, change the paradigm with which advertisers use keywords to target ads. More specifically, instead of managing the campaign based on the exact keywords that an advertiser input as targeting criteria (and not the canonical or broad match sets) advertisers can see and exclusively target sets of canonical keywords. FIG. 3 depicts an exemplary GUI 300 output to the advertiser via the advertiser's computing device. The GUI 300 includes a list of identified canonical IDs 310, in particular, the canonical IDs shown are "women's hats," "children's hats" and "men's hats." Continuing the example shown in FIG. 3, the canonical IDs were identified from the keywords "hats for women," "kid hatt," and "mens headware," respectively, which were input by the advertiser.

Returning to FIG. 2, the keyword management system is also configured to receive the input of additional keywords from the advertiser (210) and, as a result, repeat the previously

described steps for generating a set of canonical IDs (i.e., 204-208). Accordingly, the keyword management system 110 is configured to dynamically update the collection of canonicalized IDs such that, each time the set of keywords the user typed (or otherwise input) is adjusted, the collection of canonicalized IDs can be recalculated and updated and provided to the advertiser for inspection.

A finalized set of canonical IDs can then be recorded (212). For instance, the set can be recorded in keyword index 118 in association with the advertiser 104a's campaign. Subsequently, the set of canonical IDs can be utilized to perform targeting advertisements (214). More specifically, the keyword management system 110 can identify a respective set of individual keywords for each canonical ID associated with the advertiser's campaign and the search and auction engine 108 can utilize those individual keywords for targeting the service of advertisements on behalf of the advertiser. In other words, when the advertiser inputs a specific keyword during set-up, that specific keyword will be matched to a canonical version (for example: having the correct spelling and in singular, non-possessive format, etc.) and, at targeting time, the canonical version can be mapped to a set of keywords that includes all of the different variations of the specific keyword originally input (e.g., its synonyms, variations and the like).

Although principles for canonicalizing keywords is currently implemented as part of the broad match process of some existing search and keyword management systems, the process of automatically converting user input into a canonical form for output to the user and maintaining the user's keywords as a compiled set of canonical IDs for implementation in a targeted search system is not currently practiced. Particular implementations of the disclosed embodiments may realize one or more of a number of advantages for the advertising systems as well as the

advertisers themselves. In some existing advertising systems an advertiser's campaign is set up to only target the precise set of keywords that he/she entered. Accordingly, it can be appreciated that a strict exact matching approach can lack efficacy and efficiency for advertisers that inadvertently omit certain words from the precise list of targeting keywords and ignore potential synonyms and misspellings.

In addition or alternatively, when advertisers target precise keywords (which can include misspellings or abbreviations), some existing keyword management and search systems can process the keywords using a broad match algorithm that will expand the set of specific words, and the expanded set of keywords can be used for targeting and serving ads. However, the broad match expanded set of keywords also carry less precision and therefore can have a lower eCPM (effective cost per thousand impressions) when calculating bids for advertisers. Moreover, this approach can also be analytically challenging. It can be appreciated that comparing two expanded sets of keywords that result from applying a broad match algorithm to respective keywords that are related can be difficult and inefficient. In particular, there are differences in the individual keywords within the sets and the broad match algorithm will create significant overlap between the two sets. Further, trying to compare sets of keywords generated from applying a broad match algorithm is also computationally intensive because the broad match algorithm generally produces a large number of different matching keywords.

Conversely, through canonicalization pursuant to the exemplary embodiments, a single canonical keyword is identified for each keyword input by the advertiser and any repeated canonical IDs are discarded, thus, reducing redundancy of keywords and the processing that is required when selecting and serving targeted ads based on the canonicalized keywords. The disclosed embodiments can further provide significant utility and benefit for selling advertising

keywords. The disclosed embodiments are useful to advertising platforms because the canonicalization of keywords in the user input will enable the advertising platform to use broad match across all ad campaigns. Broad match expands the set of keywords that each advertisement targets which fills inventory and enables better ad targeting.

The disclosed embodiments are also valuable to advertisers because the advertisers do not have to build dictionaries of misspellings and synonyms and the targeting criteria for an advertisement can be expressed in a fraction of the size (i.e., using which makes it human readable. Most importantly advertisers can see the exact buckets of keywords that they are targeting by being presented with the corresponding canonical IDs. The user-facing features of the disclosed systems and methods fundamentally alter the methods and expectations that advertisers have for targeting advertisements and, in particular, managing keywords. By identifying the canonical IDs in advance of execution with the targeting advertising system and providing the canonical IDs to the advertisers, the disclosed systems and methods provide additional transparency for the advertiser by exposing the targeting models. In addition, management of canonical IDs instead of lists of keywords require less work on behalf of the advertiser to manage the advertising campaign and produce improved results.

It should be appreciated that, although the disclosed embodiments are discussed in the context of a keyword selection and keyword targeted advertising systems, the systems and methods can be similarly applied to other advertisement targeting mechanisms. For example, the disclosed embodiments can also be similarly applied to an end user search interface. It should also be appreciated that steps and components described above can be implemented in various orders, combinations, or subsets to achieve desirable results. In addition, the described program components and systems can generally be integrated together in one or more software products.

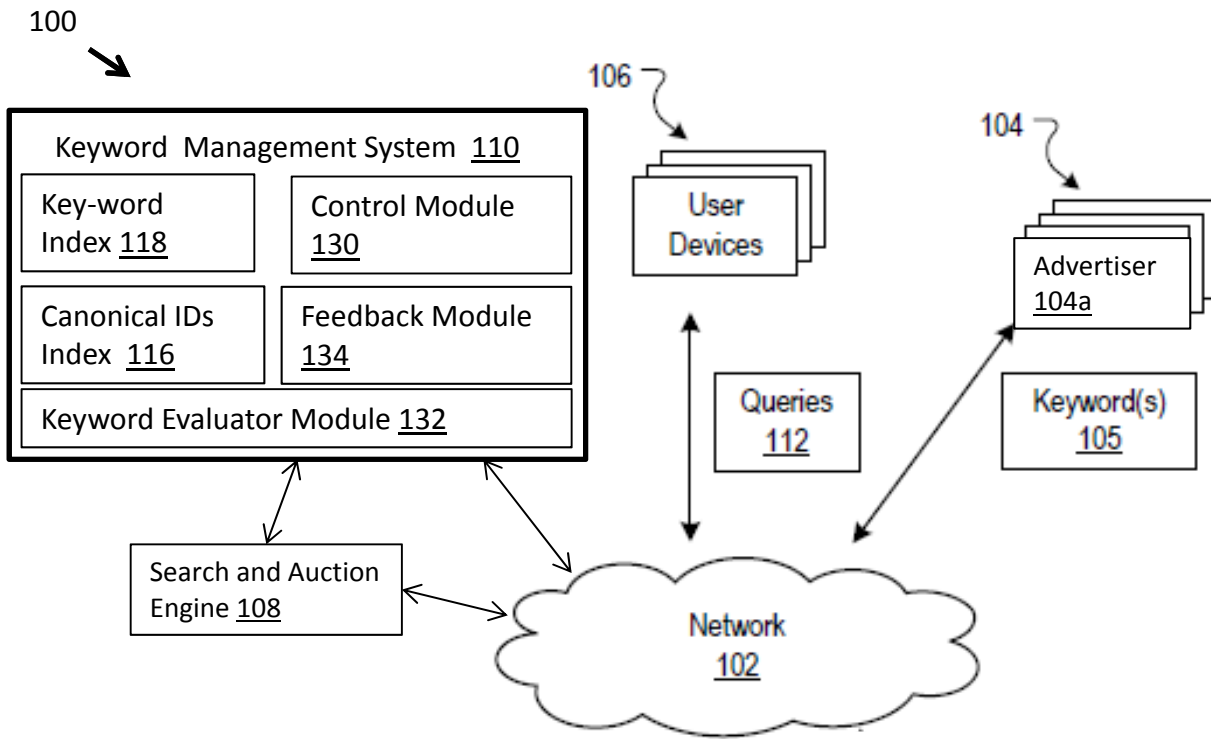


FIG. 1

Keyword Planner

Where would you like to start?

Find new keywords and get search volume data

Search for new keywords using a phrase, website or category

Get search volume data and trends

Multiply keyword lists to get new keywords

List 1

women's hats
children's hats
men's hats

List 2

Enter keywords, one per line or separated by commas to combine with the other list(s).

Targeting

- All locations
- Google
- Negative keywords

Get forecasts Get search volume

FIG. 3

200

- Identify one or more keywords for an advertising campaign (202)
- Transform the keywords into corresponding canonicalized IDs (204)
- Identify unique set of canonical IDs (206)
- Output the set of unique canonicalized keywords to the advertiser (208)
- Receive the input of additional keywords from the advertiser (210) and repeat steps (204-208)
- Record finalized set of canonical IDs (212)
- Perform targeted advertising using set of canonical IDs (214)

FIG. 2