January 06, 2016

USING MACHINE TRANSLATION FOR QUERY REWRITE GENERATION

Massimiliano Ciaramita

Anton Andryeyev

Babak Behsaz

Srinivas Narayanan

Nitin Gupta

See next page for additional authors

Inventor(s)
Massimiliano Ciaramita, Anton Andryeyev, Babak Behsaz, Srinivas Narayanan, Nitin Gupta, Neil Houlsby, Preyas Popat, and Wei Wang

USING MACHINE TRANSLATION FOR QUERY REWRITE GENERATION

ABSTRACT

A query rewrite system can be used to improve query understanding of search engines by generating multiple reformulations of the same query using machine translation. The query rewrite system receives an initial query from a user. It then retrieves known related queries to the initial query from a database. Subsequently, the system generates variant queries from the query and the related queries by passing them through a monolingual machine translator. Thereafter, the system ranks the variant queries according to predefined parameters and transmits one or more of the ranked variant queries to a search engine for further processing.

PROBLEM STATEMENT

Query rewriting is an important and integral technique utilized by search engines and question answer systems to improve the quality of returned results. Current query rewriting methods analyze search results and user clicks on the search results to identify related queries or related clusters of queries. These query rewriting methods, however, do not consider user intent at query time. Further, longer queries also suffer low recall in web search due to conjunctive term matching. There are opportunities to approach query rewriting differently to improve results for users.

DETAILED DESCRIPTION

The systems and techniques described in this disclosure relate to a query rewrite system. The system can be implemented for use via the Internet, an intranet, or another client and server environment. The system can be implemented as program instructions stored locally on a client device or implemented across a client device and server environment. The system of one or more

computers can be configured to perform particular operations or actions by virtue of having software, firmware, hardware, or a combination of them installed on the system that in operation causes or cause the system to perform the actions.

Fig. 1 illustrates an example method 100 for generating reformulations of a query using the query rewrite system.

The system receives 110 an initial query. A user can interact with the system through a client device. For example, the client device can be a computer coupled to the system via Internet. Alternatively, the system and the client device can be one machine. For example, the user can install an application on the client device which implements the system. The query received by the system can be a combination of words and alphanumeric characters or a single word. For example, the user might input "How hot is the center of the Earth?" as a query into the system. This query can be made by the user using different methods. For example, the user can submit the query by speaking the query through an audio input device associated with the client device such as a microphone. Additionally, or alternatively, the user can interact with a user interface associated with the client device to submit the query. The user interface can be any input device such as keyboard, mouse, touch screen display, etc that allows the user to enter the query.

Subsequently, the system retrieves 120 known related queries to the initial query. After receiving the query from the user, the system can access a related queries database. The related queries database can be linked to multiple query rewrite systems and search engines. All the queries made through the different query rewrite systems implemented across the world and the search engine can be indexed and stored in the related queries database. All these queries can be stored anonymously i.e., none of the queries made by different users can be linked back to the

users who made the query. The related queries database indexes all the received queries and stores them in order to provide relevant sets of related queries to the system in response to receiving an initial query.

For example, the user might submit the query “How hot is the center of the Earth”. The system would access the related queries database and look for the set of related queries. As an example, the system can retrieve historical, similar queries such as “What is the temperature at the center of the Earth”, “How hot is the Earth’s core”, “What’s the temperature of Earth’s core”, “All about Earth’s core”, etc. All of the queries retrieved by the system may not express the same meaning, see the latter example “All about Earth’s core”, but such a related query may still be important in order to improve the overall quality of the system. This is achieved by increasing the coverage of the related queries which can be related to the initial query entered in the system. Alternatively, it may be possible that the system does not contain any set of related queries for the entered query. However, the system can still work for generating reformulations of the query.

The system generates 130 variant queries from the query and the related queries. The system passes the entered query and the retrieved set of related queries to a translation system. The translation system is trained to generate a list of variants or paraphrases of an input query. The translation system can generate several translations of the entered query. For example, when the user enters the query “How hot is the center of the Earth?”, the translation system can generate various monolingual translations such as “How much hot is it at the center of the Earth”, “Is it very hot at the center of the Earth?”, etc. The translation system can generate queries which may or may not be the same as the retrieved set of queries. The translation system would generate a list of variant queries from the entered query and the set of retrieved set of

related queries. The translation system can be implemented separately from the query rewrite system or can be implemented within the query rewrite system.

After the system has generated the list of variant queries, the system ranks 140 the variant queries. The system can use any known means of ranking different queries such as by a voting mechanism which can be performed by a machine learning algorithm implemented within the system. Alternatively, the system can rank the queries by considering the resources that a search engine identifies with a query i.e., the system can rank the variant queries that are responsive to the query by: determining one or more signals for the variant query and the query, determining a score for each of the variant queries, and then ranking the variant queries based on the received scores. Examples of the signals include signals indicating relevance of the resource to the query and the signals indicating the quality of the resource.

After the system has ranked the variant queries, the system transmits 150 one or more of the ranked variant queries to a search engine. These queries can be used by the search engine to improve the ability of the search engine in providing relevant answers by increasing the coverage i.e., in case the search engine is able to respond to one form of the reformulated query but not the original query posted by the user. It can also be used by the search engine in order to improve the quality by promoting as many possible answers for all the related queries.

FIG. 2 is a block diagram of an exemplary environment that shows components of a system for implementing the techniques described in this disclosure. The environment includes client devices 210, servers 230, related queries database 250, and network 240. Network 240 connects client devices 210 to servers 230 and/or a related queries database 250. Client device 210 is an electronic device. Client device 210 may be capable of requesting and receiving data/communications over network 240. Example client devices 210 are personal computers

(e.g., laptops), mobile communication devices, (e.g. smartphones, tablet computing devices), set-top boxes, game-consoles, embedded systems, and other devices 210' that can send and receive data/communications over network 240. Client device 210 may execute an application, such as a web browser 212 or 214 or a native application 216 or a search engine application. Web applications 213 and 215 may be displayed via a web browser 212 or 214. Server 230 may be a web server capable of sending, receiving and storing web pages 232. Web page(s) 232 may be stored on or accessible via server 230. Web page(s) 232 may be associated with web application 213 or 215 and accessed using a web browser, e.g., 212. When accessed, webpage(s) 232 may be transmitted and displayed on a client device, e.g., 210 or 210'. Resources 218 and 218' are resources available to the client device 210 and/or applications thereon, or server(s) 230 and/or web page(s) accessible therefrom, respectively. Resources 218' may be, for example, memory or storage resources; a text, image, video, audio, JavaScript, CSS, or other file or object; or other relevant resources. Network 240 may be any network or combination of networks that can carry data communication.

The subject matter described herein can be implemented in software and/or hardware (for example, computers, circuits, or processors). The subject matter can be implemented on a single device or across multiple devices (for example, a client device and a server device). Devices implementing the subject matter can be connected through a wired and/or wireless network. Such devices can receive inputs from a user (for example, from a mouse, keyboard, or touchscreen) and produce an output to a user (for example, through a display and/or a speaker). Specific examples disclosed are provided for illustrative purposes and do not limit the scope of the disclosure.

DRAWINGS

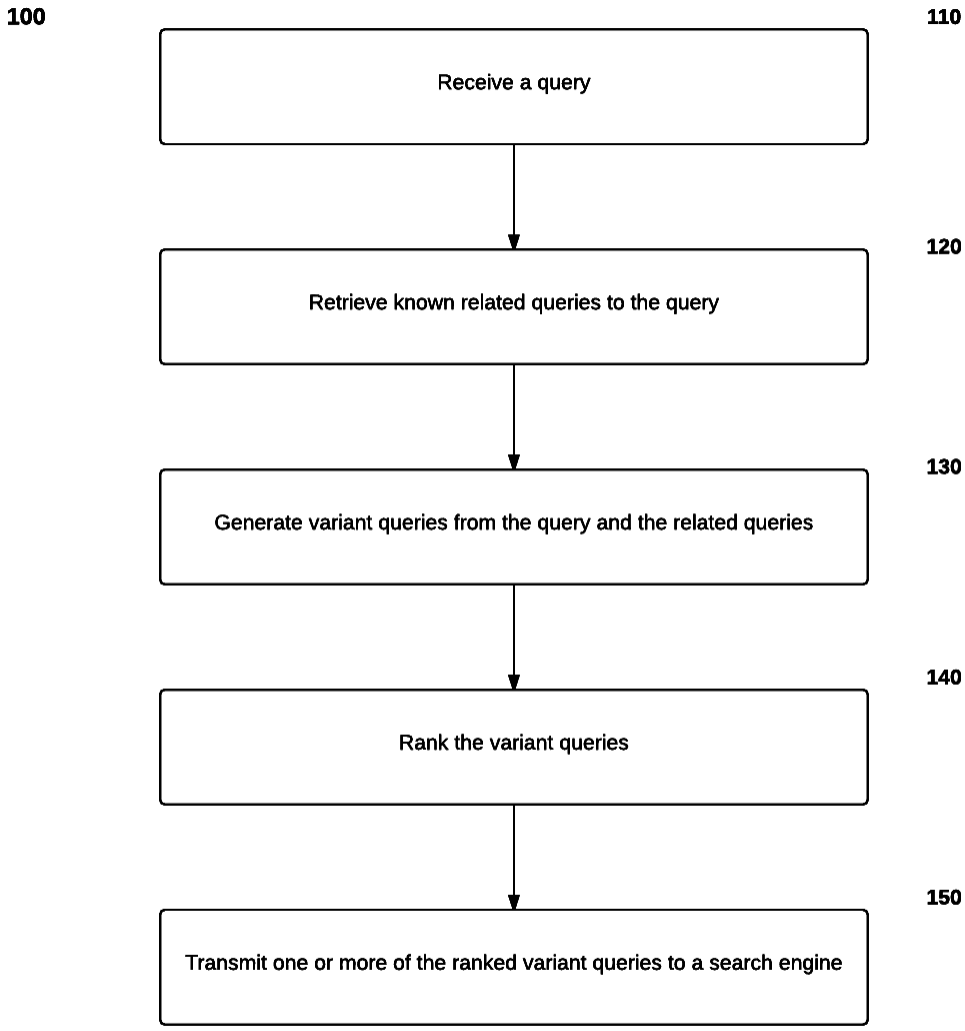

Fig. 1

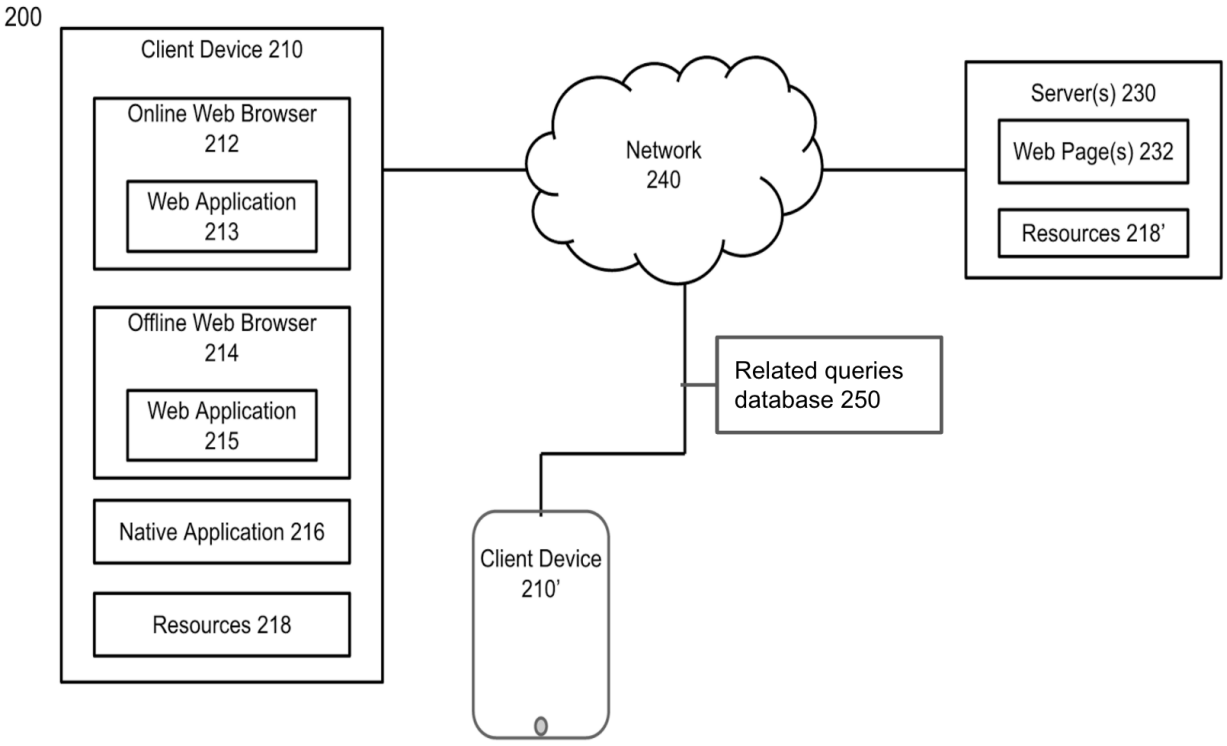
200
Client Device 210
Online Web Browser 212
Web Application 213
Offline Web Browser 214
Web Application 215
Native Application 216
Resources 218
Network 240
Server(s) 230
Web Page(s) 232
Resources 218'
Related queries database 250
Client Device 210'

Fig. 2