نظام ارجاع المعلومات المعتمد على الوكيل البرمجي

بسمة بنت صالح السلمي

المشرف الرئيسي على الرسالة

د. فتحي البرعي عيسى

د. ميسون فؤاد محمدنور ابوالخير

المستخلص

في عصرنا الحاضر أصبح الويب أكبر مستودع للبيانات، فمن الطبيعي أن يتم استخراج المعلومات من الويب. محركات البحث على شبكة الإنترنت أصحبت واحدة من أكثر الأدوات استخدامًا لاستخراج المعلومات من الويب. محركات البحث تسمح للمستخدمين البحث واسترجاع المعلومات بسهوله من خلال إدخال الاستعلام سواء كان كلمة او جمله معينة. بالرغم من أن محركات البحث تقوم بعمل جيد في البحث للعثور على صفحات معينة إلا أنها قد تكون أقل فعاليه لتلبية استفسارات واسعة أو غامضة، هذا بسبب وجود نتائج لمواضيع مختلفة ومعاني عديدة لنفس الاستعلام مما يُلزم المستخدم للبحث في عدد كبير النتائج الغير متعلقة في موضوعه حتى يجد النتيجة التي يريدها. من جانب آخر هنالك عدد كبير من التكرار في نتائج البحث مما يعكس سلبا ويزيد من بحث المستخدم للحصول على النتيجة المرجوة.

هذه الرسالة تعالج مشكلتين أساسيتين هما التكرار وغموض المعنى للكلمات. حيث أن تعدد المعاني وتكرار النتائج سبب من أسباب ضعف أداء نظام استرجاع المعلومات. التكرار في صفحات الويب (Near Duplicate webpages) ستحل عن طريق أداة تقوم بكشفها وإزالتها حيث ان إزالة هذا التكرار يعود بالنفع في اوجه كثيرة منها تقليل عدد النتائج البحث هذا يؤدي إلى تقليل وقت البحث مما يسمح للمستخدم بأن يجد رغباته بأسرع ما يمكن. تنظيم نتائج البحث في مجموعات (Search result Clustering) من طرق حل غموض الكلمات وتعدد معانيها حيث أنه يقوم بتنظيم نتائج البحث إلى مجموعة من المجموعات ذا المعنى الواحد. تصميم خوارزميه لهذا التنظيم امر ليس بسهل لأنه يجب علينا أن نتأكد أن الاسم المختار لكل مجموعة مقروء وذو معنى ويمثل هذه المجوعة بالفعل.

تقترح هذه الرسالة بناء نظام استرجاع المعلومات المعتمد على الوكيل البرمجي يقوم بعمليه تصفيه النتائج من التكرار و تجميع البيانات وتنظيمها عن طريق اضافة مكوني التجميع والتصفية (Filtering and Clustering Component). وكلاء مكون التجميع (Agent Based Clustering Component) يتعاملوا مع صفحات الويب من خلال تجمعها وتنظيمها في مجموعات بناء على المعنى الفعلي للكلمة. الورد نت (WordNet) يستخدم للمساعدة في عمليه التنظيم بناء على مترادفات WordNet. وكلاء مكون التصفية (Agent Based Filtering Component) يقوموا بإزالة التكرار من نتائج البحث. أثبتت التجارب أن مكون التصفية المقترح يحقق دقه (precision) بنسبة تصل الى 97 % و استدعاء (recall) يصل الى 97 % و مكون التجميع حقق دقه في معنى اسم المجوعة (cluster label) تصل الى 92 % .

Agent Based Information Retrieval System

Bassma Saleh Alsulami

Prof Thesis Advisor

.Dr. Fathy A. Eassa

DR. Maysoon F. Abulkhair

Abstract

The Web has become the largest easy available repository of data. Hence, it is natural to extract information from it and Web search engines have become one of the most used tools in Internet. Search engines allow the user to search and retrieve information in simple and easy way using terms such as phrase or keyword. Search engines retrieve web pages from its database that match the search terms entered by the searcher. However, while search engines are definitely good for certain search tasks such as finding the home page of an organization, they may be less effective for satisfying broad or ambiguous queries. The results on different subtopics or meanings of a query will be mixed together in the list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is a lot of the duplicated or near duplicated webpages in the search results.

This thesis addressed and solved the two main problems: Near duplication and Word sense ambiguities (multiple meaning). Word ambiguity and a lot of near duplicate may lead to poor performance in Information Retrieval (IR) systems. Near Duplications Document (NDD) detection is used to solve duplication problem. Removing the NDD has a lot of advantages, it reduces search result list that leads to decrease search time that allow the user find her/his requirement as fast as possible. Search results clustering is an attempt to solve multiple meaning problems by automatic organizing the linear lists of document references returned by a search engine into a set of meaningful thematic categories. Designing a Web search clustering algorithm is a big challenge because readable and unambiguous labels of the thematic groups are an important factor of the overall quality of clustering; also labels should be a good descriptive to the search result cluster.

Multi-agent based information retrieval system is proposed to enhance the search process by adding clustering and filtering components. Cluster agent based component was added where agents deal with the web search engine results by clustering them into relevant synonym's category for given queries. WordNet was employed to classify results in search engine result page in appropriate synonym's category according to WordNet synsets. Filtering Agent based component was embedded to eliminate the near duplicate data references. The experiments show that proposed filtering component leads to a precision of over 96%, and a recall of over 97% and the proposed clustering component leads to quality of cluster’s title over 92%.