بخشی از ترجمه مقاله

عنوان فارسی مقاله :

پلت فورم مبتنی بر هدوپ برای پردازش زبان طبیعی اسناد و صفحات وب

عنوان انگلیسی مقاله :

A hadoop based platform for natural language processing of web pages and documents

توجه !

این فایل تنها قسمتی از ترجمه میباشد. برای تهیه مقاله ترجمه شده کامل با فرمت ورد (قابل ویرایش) همراه با نسخه انگلیسی مقاله، اینجا کلیک نمایید.

بخشی از ترجمه مقاله

5. Conclusions and future work

In this paper, a distributed system for crawling web documents and extracting keywords and keyphrases has been presented. The parallel architecture is provided by implementing the Apache Hadoop platform, while text annotation and key features extraction rely on the NLP opens source GATE platform. The main contributions offered by our work is the capability of executing general purpose GATE applications (including a wide range of NLP activities) in a distributed design (exploiting the benefits of scaling performances, especially for very large text corpora) with minimal code update and without the need for programmers to care about parallel computing constraints, such as task decomposition, mapping and synchronization issues. Evaluating processing performances on different cluster configurations (from 2 to 5 nodes) has showed a nearly linear scalability of the system, which is an encouraging result for future assessments on even larger datasets and cluster configurations. These actually represents open issues for future work. Moreover, it could be interesting to implement our keywords/keyphrases extraction module on other parallel computing environment, such as the cited Spark. Furthermore, in order to improve the quality of key features extraction, external knowledge resources could be used, especially Semantic repositories and frameworks, to allow the annotation of semantic features and relations.

۵. نتیجه گیری و برنامه های آینده

در این مقاله ، سیستم توزیعی برای کرالینگ اسناد وب و استخراج کلیواژه ها و عبارات ارائه شده است.

ساختار موازی با اجرای پایگاه هدوپ آپاچی ارائه شده در حالی که تفسیر متن و استخراج ویژگی های کلیدی متکی بر پایگاه گیت منبع باز NLP است. سهم اصلی پیشنهادی کار ما ، قابلیت اجرای کاربردهای گیت چندمنظوره (شامل دامنه ی وسیعی از فعالیت های NLP) در طراحی توزیع شده (با استفاده از مزیت عملکردهای مقیاس بندی ، خصوصا برای مجموعه نوشتاری متنی بسیار حجیم) با به روز رسانی کد مینیمم و بدون نیاز به برنامه نویس برای نظارت بر محاسبه ی موازی محدودیت هایی مثل تجزیه ی متن ، نقشه کشی و مسائل همگام سازی است. ارزیابی عملکردهای پردازش روی اشکال مختلف خوشه ای (از ۲-۵ گره) قابلیت مقیاس پذیری تقریبا خطی سیستم را از خود نشان داده که نتیجه ی مطلوب ارزیابی های آینده در ترکیب بندی های ی خوشه ای و مجموعه داده های بزرگتر است . که در واقع نشان دهنده ی مسائل باز به روی کارهای آینده است. علاوه براین ، اجرای مدول استخراج کلیدواژه/عبارت اصلی روی محیط محاسباتی موازی دیگر ، مثل اسپارک ، کار جالبی است. علاوه براین ، برای ارتقای کیفی استخراج ویژگی های اصلی ، منابع دانش خارجی را می توان به کار برد، خصوصا چارجوب ها و مخازن معنایی ، تا تفسیر ویژگی های معنایی و ارتباط آنها فراهم شود.