قسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة عين شمس

تحسين عملية استنتاج العلامات النصية من الوثائق اعتمادا على المعنى الدلالى

رسالة مقدمة للحصول على درجة الماجيستير في الحاسبات والمعلومات

إعداد

ايمان اسماعيل سيد

بكالوريوس الحاسبات والمعلومات
معيدة بقسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة عين شمس

تحت إشراف

أ.د. خالد البهنسي
أستاذ بقسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة عين شمس

د. محمد العليمى
أستاذ مساعد بقسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة عين شمس

د. ولاء جاد
أستاذ مساعد بقسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة عين شمس

ملخص الرسالة

في الآونة الأخيرة، كمية البيانات المتوفرة على شبكة الإنترنت في تزايد سريع، لذلك فمن الصعب البحث عن البيانات ذات الصلة في مجموعة بيانات ضخمة. وضع علامات نصيه للمستند يوفر حلا لهذا النوع من المشاكل. حيث انها تعطي النص معلومات إضافية في شكل ملاحظات أو تعليقات. التعليق التوضيحى للمستند يسهل مهمة العثور على الموضوعات الرئيسية للوثيقة . علاوة على ذلك، فأنه يساعد القارئ على القاء نظرة عامة على المستند وفهمه.

بسبب انتشار تطبيقات وسائل الاعلام الاجتماعية مثل الفيسبوك، تويتر ... الخ. يتم إنتاج الملايين من المستندات القصيرة يوميا. لذلك، يتم استخدام تصنيف النص لاكتشاف المعرفة من هذه المستندات النصية غيرالمنظمه. المستندات النصية القصيرة لها خصائص خاصة حيث انها مليئة بالاخطاء وتحتوى على كلمات قليله لذلك كلماتهم نادرا ما تتكرر. وتستند الأساليب التقليدية لتصنيف هذه الأنواع من الوثائق إلى طريقة "BOW" التي تصنف الوثائق النصية ككلمات مستقلة .. بحيث يعبر عن كل مستند بمجموعه من الكلمات' كل كلمه لها قيمه مساويه لعدد مرات ظهورها في المستند. . فاستخدام الأساليب التقليدية للتصنيف كمجموعه من الكلمات لديها العديد من العيوب: المستندات النصيه القصيره ليس لديها عدد كاف من الكلمات لتمثيل المستندات باستخدام Bow حيث أن هذه الطريقه تعتمد على عدد تكرار الكلمات فى المستند، والكلمات فى المستندات النصيه القصيره نادره التكرار. ايضا طريقه BOW لم تاخد فى الاعتبار العلاقه بين الكلمات وبعضها . لذا يستخدم المعنى الدلالى فى وضع العلامات النصيه للتغلب على المشاكل الموجوده فى المستندات النصيه القصيره. ولاثراء وتزويد المستندات ببعض المعلومات الاضافيه لفهم المستند اكثر وتصنيفها جيدا.

فى هذا العمل ، تم اقتراح نموذجان فعالان لاستنتاج العلامات النصيه من المستندات اعتمالا على المعنى الدلالى .النموذج الاول CBER معتمد على اثراء علميه التصنيف للمستندات النصيه حيث انه مكون من النموذج المقترح SAWN والذى يعتمد على قاعده البيانات (WordNet) لاستنتاج المعنى الدلالى ، والنموذج الاخر WVTFالذى يعتمد على BOWفى تمثيل البيانات النصيه لاداء عمليه التصنيف. النموذج SAWN يستخراج الكلمات ذات المعنى من المستندات النصيه القصيره باستخدام قاعده البيانات (WordNet) ويعطيها قيمه تعبر عن اهيمتها بالاستعانه بمرادفات هذه الكلمات . وبالتالى الكلمات التى لها نفس المعنى تزيد من قيمه مرادفاتها . علاوه على ذلك تم استغلال العلاقه بين الكلمات وبعضها فى حل المشاكل المتعلقه بالمعنى الدلالى مثل تواجد كلمه لها اكثر من معنى او العكس بان توجد مجموعه من الكلمات تعبر عن نفس المعنى.

النموذج المقترح CBER يعمل على اثراء المستندات النصيه القصيره بمعلومات اضافيه دون الحاجه لزياده عدد الكلمات المعبره عن كل مستند .كما انه ياخد فى الاعتبار الكلمات التى لا توجد فى قاموس ال (WordNet) فانه يعطى للكلمات قيمه للتعبير عن اهميتها اعتمادا على قيمتها الاساسيه التى تعتمد على عدد مرات تكرار الكلمه و قيمه اخرى اعتمادا على المعنى الدلالى ، لذا تاخد الكلمات ذات المعنى والمعرفه فى القاموس قيمه اكبر عن غيرها من الكلمات الغير معرفه .

النموذج ا الثانى "Wiki_Spots" يستخدم للتعبير عن المعنى الدلالى باستخدام الويكيبيدا عن طريق ربط المستندات النصيه القصيره بالويكيبيدا لاستخراج الكلمات التى تستخدم للاشاره عن مقالات توضيحيه للمعنى الدلالى لهذه الكلمات ، والتعبير عن المستندات بطريقه جديده باستخدام هذه الكلمات التى تعبر عن الموضوعات المتعلقه بالمستندات ، ومن ثم يتم اعطاء هذه الكلمات(Spots) قيمه اعتمادا على اهميتها فى المستند للتصنيف الجيد. علاوه على ذلك ، فان كل spot يمكن ان يعبر عن كلمه واحده او جمله اسميه قصيره مما يحسن من نتائج عمليه التصنيف.

واظهرت التجارب العمليه ان نموذج CBER لاثراء عمليه التصنيف جيد فى تصنيف المستندات النصيره القصيره الذى وصل الى ٠,٩٣٣ ، ٠,٩٣٦ ، ٠,٩٣٤ و٩٤% فى الدقه والتذكير ، ونموذج SAWN وصل إلى ٠,٨٧٢، ٠,٨٧٣، ٠,٨٧٧ و ٨٨٪. أيضا، نموذج ويكيبيديا القائم على "Wiki_Spots" وصل إلى ٠,٨٣٢ ، ٠,٨٣٣، ٠,٨٣٧ و ٨٤٪ بالمقارنة مع طريقة BOW التقليدية.

Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University

An Enhanced Automatic Model Based On Semantic Annotation For Text Documents

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of science in Computer and Information Sciences.

By
Eman Ismail Sayed

B.Sc. in Computer and Information Sciences (2012),
Demonstrator at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Khaled El-Bahnasy

Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

Dr. Mohamed Eleliemy

Associate Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

Dr. Walaa Gad

Associate Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

Cairo - 2017

Abstract

The amount of available data on the web is rapidly increasing, so it is difficult and more complicated to search and identify relevant data in huge data sets. Text document annotation provides one approach to ease such types of problems. It provides the text with additional information in the form of notes or comments. Annotation facilitates the task of finding the main topics of a document. Moreover, annotation helps the reader to overview and understand the document.

Due to the spread of social media applications such as Facebook, Twitter.. etc, millions of short texts are being produced daily. Text classification is used to discover knowledge from these unstructured text data. The short text documents (STDs) have special characteristics as being noisy and sparsy because their words are rarely repeated. The traditional methods of classifying such types of documents are based on the Bag of Words (BOW) method, which indexes text documents as independent features. Each feature is a single term or word in a document. A document is represented as a vector in feature space. A document vector contains the word weights, which are the number of word occurrences in the document. Classification of STDs based on BOW has many drawbacks: STDs do not provide enough co-occurrence of words or shared context. Representation of such documents is almost sparse because of empty weights

when using BOW. Hence, the traditional bag of words (BOW) method fails to achieve high accuracy. Moreover, the BOW method treats synonymous words as different features and does not consider the relations between words and documents. Therefore, semantic knowledge was introduced as a background to focus on the semantic relationships between the documents words or terms.

In this thesis, two effective models for semantic annotation are proposed. The first model is Classification Based on Enrichment Representation (CBER). It is composed of the proposed semantic analysis based on WordNet(SAWN) model and the word vector term frequency (WVTF). WVTF is a BOW representation of text documents. SAWN maps the text documents with WordNet to extract the concepts. Concepts are the terms that are defined in WordNet. SAWN chooses the most suitable synonyms for s document concepts by studying and understanding the surrounding concepts in the same document. Thus, concepts with the same meaning will increase the weight of their synonyms. Furthermore, the semantic relationships between concepts have been exploited in order to solve the disambiguation problems such as polysemy and synonyms.

The CBER model enriches the STDs with semantic weights to solve disambiguation problems without increase the document features. It considers all documents terms. However, some terms may not be defined in WordNet. The terms are then provided with a new weight. These weights depend on the term frequency weight based on WVTF and the semantic weight based on SAWN. Thus, the terms that are defined in WordNet gain more weight.

The second model is the Wiki_Spots model. It identifies and extracts cross-referencing text (spot) from the documents using the Wikipedia knowledge base. Each spot is annotated to a Wikipedia article (page) considering the relationships with other spots in the same document. Wiki_Spots model exploits these spots to represent the documents as vectors of topics rather than vector of words in the traditional BOW method. Moreover, each spot is unigram or a noun phrase that helps increasing the classification accuracy.

The experimental results showed that the proposed CBER model is valuable in annotating short text documents to their best labels. CBER showed significant performance 0.933, 0.936, 0.934 and 0.94 in precision, recall, F-measure and accuracy respectively, and SAWN reached 0.872, 0.873, 0.877 and 0.88. Also, the Wikipedia based model "Wiki_Spots" reached 0.832, 0.833, 0.837 and 0.84 compared to the traditional BOW method.

Acknowledgements

Thanks first and foremost to Allah, who gives me the knowledge and patience to produce this work. Thanks Allah to response to my prayers to achieve my goal.

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Khaled Bahnsy for the continuous support of my M.sc study and related research. For his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my M.Sc study.

Also, I would like to extend my thanks and appreciation to my supervisor Dr. Mohamed Hamdy for his guidance, constant support and useful feedback that helped me finalizing my work and my M.Sc thesis.

I also appreciate Dr. Walaa Gad for her effort with me beginning with preparing the proposed master idea till writing and reviewing the thesis, she always followed up my work weekly and her scientific contributions helped me to advance my work and to get better results.

I would to thank all my family especially my mother who helps me in everything. Also, I am lucky to be married to Ali Gamal, who helped me to finalize writing and reviewing my thesis.

List of publications

- Eman Ismail and Walaa Gad. CBER: An Effective Classification Approach Based on Enrichment Representation for Short Text Documents. Journal of Intelligent Systems. Vol.26, pp. 233-241. 2016.
- Eman Ismail, Walaa Gad, Mohamed Hamdy, and Khaled Bahnsy. Text document annotation methods. In 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 634-640. IEEE, 2015.

Contents

List of Figures

List of Tables

Abbreviations

BOC	**B**ag **O**f **C**oncepts
BOW	**B**ag **O**f **W**ords
CAT	**C**ontent **A**nalysis **T**oolkit
CFP	**C**all **F**or **P**apers
CBER	**C**lassification **B**ased **E**nrichment **R**epresentation
DB	**D**ata **B**ase
DW	**D**escriptive **W**ords
IR	**I**nformation **R**etrieval
KW_SA	**K**eyword **S**emantic **B**ased **A**nnotation
KW_OA	**K**eyword **O**ntolgy **B**ased **A**nnotation
KNN	**K** **N**earest **N**eighbour
KWBA	**K**eyword **B**ased **A**nnotations
LDA	**L**atent **D**irichlet **A**llocation
LSA	**L**atent **S**emantic **I**ndexing
LDA	**L**atent **D**irichlet **A**llocation
LSA	**L**atent **S**emantic **I**ndexing
NLP	**N**atural **L**anguage **P**rocessing
NGD	**N**ormalized **G**oogle **D**istance
NBC	**N**aive **B**ayes **C**lassifier
OBA	**O**ntology **B**ased **A**nnotation
OWL	**O**nline **W**riting **L**ap
POS	**P**art **O**f **S**peech