International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue14
November 2017

Short Texts Analasisby Using Semantic Enrichment and Hashing

Dimple Priya & P . Pradeep
Sir Vishveshwaraiah Institute Of Science And Technology
M.Tech, Assistant Professor Sir Vishveshwaraiah Institute Of Science And Technology
Pristinezzzz@Gmail.Com , Pradeep.Pathi89@Gmail.Com

ABSTRACT:

Clustering short texts by their meaning is a challenging task. The semantic hashing approach encodes the meaning of a text into a compact binary code. Thus, to tell if two texts have similar meanings, we only need to check if they have similar codes. The encoding is created by a deep neural network, which is trained on texts represented by word-count vectors (bag-of-word representation). Unfortunately, for short texts such as search queries, tweets, or news titles, such representations are insufficient to capture the underlying semantics. To cluster short texts by their meanings, we propose to add more semantic signals to short texts. Specifically, for each term in a short text, we obtain its concepts and co-occurring terms from a probabilistic knowledge base to enrich the short text. Furthermore, we introduce a simplified deep learning network consisting of a 3-layer stacked auto-encoders for semantic hashing. Comprehensive experiments show that, with more semantic signals, our simplified deep learning model is able to capture the semantics of short texts, which enables a variety of applications including short text retrieval, classification, and general purpose text processing

INTRODUCTION

With explosive growth of mobile devices including smart phones, PDAs, and tablet computers and the applications installed in them, the mobile-Internet will maintain the development growth trend as 4G communication network is extensively promoted to our lives. What users of the mobile devices and applications need is that mobile-Internetcanprovide them with the service which is user-friendly, highspeed, and steady. In addition, the security issues of mobile terminals and the Internet access are attached importance to. And as a combination ofcloucomputing, mobile devices and wireless networks, mobile cloud computing is an emerging
but very promising paradigm which brings rich computational resources to mobile users, network operators, as well as cloud computing providers The flaws of data storing and data computing in mobile-Internet applications can be overcome by mobile cloud computing while the new paradigm can also accomplish cloud based multi-user data sharing, end
Geographical service limitation and process real-time tasks efficiently at the same time. There is no accurate definition of mobile cloud computing, several concepts were proposed, and two most popular schemes can be described as follows:
Mobile cloud computing is a kind of scheme which could run an application such as a weather monitor applicationon remote cloud servers as displayed in Figure, whileThe mobile devices just act like normal PCs except that the mobile devices connect to cloud servers via 3G or4G while PCs through Internet. And this concept is computing Taking advantages of leisure resources such as CPU, memory, and storing disks, another model of mobile cloud computing exploits the mobile devices themselves as resources providers of cloud. And the scheme supports user mobility, and recognizes the potential of mobile clouds to do collective sensing as well. In this paper, we mainly use the first paradigm mentioned above, but the second one inspires us to assume that what if the mobile devices do not provide computing resources or storing resources but sensing data instead? In fact, most mobile devices are capable to capture some data from the environment nowadays, for example, almost Every smart phone are equipped with sensors of proximity, accelerometer, gyroscope, compass, barometer, camera, GPS, microphone, etc. Combining the concept of WSN, mobile devices can be regarded as mobile sensors that are able to provide other mobile devices who are users of the mobile cloud services with some sensing information including environment monitoring data, health monitoring data, and so on. We take a weather monitor application as an example in this paper. Assuming that a company develops a weather monitor application which aims to share real-time weather information such as temperature, humidity, pictures, and precise location information and so on to other users of the application. And the application utilizes the user-cloud-user model instead of peer to-peer model so that the users can get classified and demanded information. Another feature of the application is that the users are divided into different hierarchies, depending on which users can get different sensing data, and users with higher privilege level can, of course, get access to more specific and more frequently updated information. In order to meet what the application requires, security issues of the whole system should not be ignored, among all security issues the most important two security issues in such model can be divided into two parts: authority of application

MODULES:

Register This is sub module of User Module if here only user first register for the account. Login This is also sub module of user here user after registration user can login. Search Extensive research studies have been done on structured queries as well as on text search over short keyword queries. In the view of difficulty of formulating the queries with precise structures over standard data, an IR-style querying, in particular, full text and keyword search is introduced. This approach has the merit of eliminating structures in the query. Maio.et.al presented an ontology based retrieval

approach, which supports data organization and visualization and provides a friendly navigation model. Top k Search This is also sub module of user module here admin added one URL. That URL search top k search. So here that URL search tops more by the user and first position. And algorithm also using top k search for the searching keyword on top position.

Admin

Login This is sub module of Admin module and here only admin can login This default password for the admin. Admin handle the user details and modification show the all details of user. All Users In this module show all user list who is searching the top k search This module used based on the algorithm this is main module of this project because. Add searches link Here in this module admin added link for the user so user searches on top position. That link show on top position that's why this is depends on the algorithm. View Uploaded Here admin show all uploaded the files and view the files of users also this is only module modified by the admin not user this restricted for the users. Most Searches Here admin add some link that link search by the users so if user search more time any link so rank also decide for that link which more then search by the users . so this is most searches link by the users.

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

OBJECTIVES

process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

LITERATURE SURVEY:

In this brief, we propose a new method to reduce the number of support vectors of support vector machine (SVM) classifiers. We formulate the approximation of an SVM solution as a classification problem that is separable in the feature space. Due to the reparability, the hard-margin SVM can be used to solve it. This approach, which we call the separable case approximation (SCA), is very similar to the cross-training algorithm explained in , which is inspired by editing algorithms . The norm of the weight vector achieved by SCA can, however, become arbitrarily large. For that reason, we propose an algorithm, called the smoothed SCA (SSCA), that additionally upper-bounds the weight vector of the pruned solution and, for the commonly used kernels, reduces the number of support vectors even more. The lower the chosen upper bound, the larger this extra reduction becomes. Upper-bounding the weight vector is important because it ensures numerical stability, reduces the time to find the pruned solution, and avoids over fitting during the approximation phase. On the examined datasets, SSCA drastically reduces the number of support vectors. In this paper a multi-level fuzzy min-max neural network classifier (MLF), which is a supervised learning method, is described. MLF uses basic concepts of the fuzzy min-max (FMM) method in a multi-level structure to classify patterns. This method uses separate classifiers with smaller hyperboles in different levels to classify the samples that are located in overlapping regions. The final output of the network is formed by combining the outputs of these classifiers. MLF is capable of learning nonlinear boundaries with a single pass through the data. According to the obtained results, the MLF method, compared to the other FMM networks, has the highest performance and the lowest sensitivity to maximum size of the

EXISTING SYSTEM:

Many approaches have been proposed to facilitate short text understanding by enriching the short text. More effectively, a short text can be enriched with explicit semantic information derived

from external resources such as WorldNet, Wikipedia, the Open Directory Project (ODP), etc.Salakhutdinov and Hinton proposed a semantic hashing model based on Restricted Boltzmann Machines (RBMs) for long documents, and the experiments showed that their model achieved comparable accuracy with the traditional methods, including Latent Semantic Analysis (LSA) and TF-IDF.

DISADVANTAGES OF EXISTING SYSTEM:

Search-based approaches may work well for so-called head queries, but for tail or unpopular queries, it is very likely that some of the top search results are irrelevant, which means the enriched short text is likely to contain a lot of noise. On the other hand, methods based on external resources are constrained by the coverage of these resources. Take WordNet for example, WordNet does not contain information for proper nouns, which prevents it to understand entities such as “USA” or “IBM.” For ordinary words such as “cat”, Word Net contains detailed information about its various senses. However, much of the knowledge is of linguistic value, and is rarely.

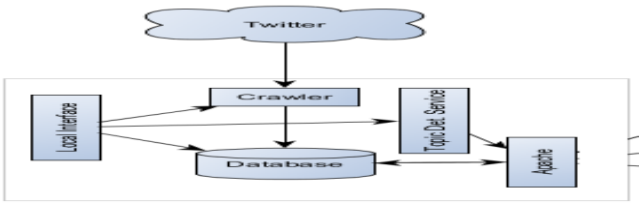
PROPOSED SYSTEM:

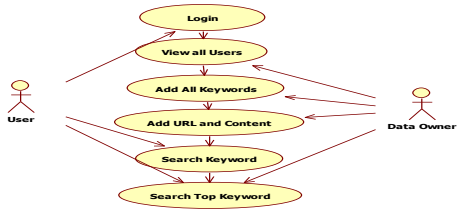
In this paper, we propose a novel approach for understanding short texts. Our approach A semantic network based approach for enriching a short text; We present a novel mechanism to semantically enrich short texts with both concepts and co-occurring terms, such external knowledges are inferred from a large scale probabilistic knowledge base using our proposed thorough methods. For each auto encoder we design a specific and effective

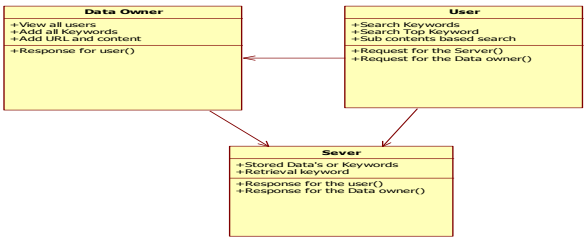
ADVANTAGES OF PROPOSED SYSTEM:

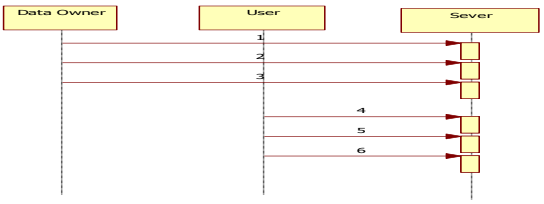
We carry out extensive experiments on tasks including information retrieval and classification for short texts. We show significant improvements over existing approaches, which confirm that Concepts and co-occurring terms effectively enrich short texts, and enable better understanding of them; our auto-encoder based DNN model is able to capture the

SYSTEM ARCHITECTURE









HARDWARE REQUIREMENTS

:System Pentium Dual Core Hard Disk 120 GB.Monitor 15’’ LED Input Devices Keyboard, MouseRam1GB

.

SOFTWARE REQUIREMENTS:

Operating system Windows 7.Coding Language JAVA/J2EEToolNetbeans 7.2.1Database MYSQL
SYSTEM STUDY The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

methods that are employed to educate the user about the system and to make him familiar

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items: Valid Input identified classes of valid input must be accepted. Invalid Input identified classes of invalid input must berejected.Functions identified functions must be exercised. Output identified classes of application outputs must be exercised. Systems/Procedures: interfacing systems or procedures must be invoked.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

All field entries must work properly.

Pages must be activated from the identified link.
The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.
- Integration Testing
-

Test Results:

All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered

HARDWARE CONFIGURATION

System Pentium IV 2.4 GHz. Hard Disk 40 GB.Monitor 15 VGA Colour.Mouse Logitech. Ram 1 GB.

SOFTWARE CONFIGURATION

Operating system Windows XP/7/8.Coding Language JAVA/J2EEIDE Eclipse Database MYSQL

CONCLUSION:

To evaluate the effectiveness of enriching short texts and DNN semantic hashing, we carried out two experiments. First, we perform an information retrieval experiment on MNS News data, and then we do a classification task on partial Wikipedia data. We compare our enrichment method with other popular knowledge-based enrichment approaches, and we also compare our DNN model with other retrieval methods, such as TF-IDF, LSA and RBMs-based semantic hashing model. The experiments clearly demonstrate the benefits of utilizing our proposed enrichment method and the deep neural network respectively, moreover, the combination of them as an unified framework significantly improve the accuracy of retrieval and classification tasks, which indicates that our enrichment method indeed introduce useful semantic signals into short texts, and the learned binary codes successfully capture the semantic similarity between texts.

REFERENCE:

D. M. Blei, A. Y. Ng, and M. I. Jordan.Latent dirichlet allocation.Journal of Machine Learning Research, 3:993–1022, 2003 E. Gabrilovich and S. Marko itch. Feature generation for text categorization using world knowledge. In IJCAI, pages 1048–1053, 2005E.

Gabrilovich and S. Marko itch.Computing semantic relatednessusingwikipediabasedexplicitsemanticanalysis.I– 1611, 2007.X. Hu, N. Sun, C. Zhang, and T.-S.Chua.Exploiting internal and external semantics for the clustering of short texts using world knowledge.In CIKM, pages 919–928, 2009.D. Kim, H.

Wang, and A. H. Oh.Context-dependent conceptualization.In IJCAI, 2013.M. Sahami and T. D. Heilman.A web-based kernel function for measuring the similarity of short text snippets. In WWW, pages 377– 386, 2006.R. Salakhutdinov and G. E. Hinton.Semantic hashing. Int. J. Approx. Reasoning 50(7):969–978, 2009