International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

Contextual Advertising through Entity Extraction

Asmita Joshi, J.S.Sodhi, Roopali Goel

Abstract— Contextual Advertising is a type of Web advertising Content match has greater potential for content providers, publishers and advertisers, because users spend most of their time on the Web on content pages. In past researches, Contextual targeting technology works by searching the website and looks up relevant keywords But Nowadays, In contextual advertising, matching is determined automatically by the page content, which complicates the task considerably. We Proposed a System which can target the large group of consumer on internet. In Our system we make contextual targeting more relevant with Extraction of relevant entities from the web page. We extract the entities from web page, which is of interest to the consumer. We target the interest of internet user and put up the ads according to their interest. The system is designed in such a way that it can extract entities (Name, Place, Title, Location, date etc) from web page and ad publisher put up a advertise on that page which include those entities which are extracted from page. This Process will extract different types of entities, which will identify by different patterns prepared by the rules based approach. The described system able to find out the entities in many context using pattern identification. Once pattern will match entities are extracted and used by ad publisher for publishing the ads according to the context of entities. The above described method is more relevant and effective and it will target more consumers and generate revenue by advertising.

Keywords -

I. INTRODUCTION

The internet has become an ongoing emerging source that tends to expand more and more. The growth of this particular medium attracts the attention of advertisers as a more productive source to bring in consumers. A contextual ad system scans the text of a website for entities and returns ads to the web page based on what the user is viewing, either through ads placed on the page or popup ads.

For example:-"If I'm on a site reading about LED televisions, they show me ads for retailers who sell them - without the publisher or the advertiser (or even the ad network) has to explicitly specify anything. This is really just contextual advertising."

Contextual advertising is a form of targeted ads serving to users of various Internet services. We can also called it 'in-text' or 'in-context' advertising Advertising touches challenging problems concerning how ads should be analyzed, and how systems accurately and efficiently select the best ads. Information retrieval systems were designed to capture "relevance", and relevance is a basic concept in advertising as well. As with document retrieval, in the context of advertising we assume that an ad that is topically related to a Web page is relevant. Elements of an ad such as text and images tend to be mutually relevant, and often ads are placed in contexts which match the product at a topical level.

A. Contextual advertising is interplay of four players

1. The publisher is the owner of the web pages on which the advertising is displayed. The publisher typically aims to maximize advertising revenue while providing a good user experience.
2. The advertiser provides the supply of ads. Usually the activity of the advertisers is organized around campaigns which are defined by a set of ads with a particular temporal and thematic goal (e.g. sale of digital cameras during the holiday season). As in traditional advertising, the goal of the advertisers can be broadly defined as the promotion of products or services.
3. The ad network is a mediator between the advertiser and the publisher and selects the ads that are put on the pages. The ad-network shares the advertisement revenue with the publisher.
4. Users visit the web pages of the publisher and interact with the ads. Contextual advertising usually falls into the category of direct marketing (as opposed to brand advertising), that is advertising whose aim is a "direct response" where the effect of a campaign is measured by the user reaction.

II. PROBLEM DEFINITION

The kind of system is need to develop which can target a large number of consumers from different interests and choices .In the current scenario of marketing many website earn revenue from this recent trend of advertising. So the system is able to identify the entities in different context by using patterns. There is need to find the efficient way of making publishing of ad's more relevant.

Contextual advertising task can be divided into different segments of subtask ,in the first phase of project We have to extract the entities from web pages e.g.(Name, organization, place etc) and Extracted entities are matched with the entities of ad pages then desired ads are displayed on the screen of websites.

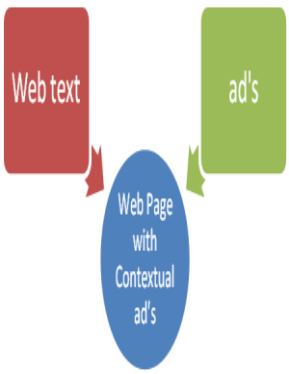

Fig 1: Contexual ad's Concept

Manuscript published on 28 February 2013.
***** Correspondence Author (s)
Asmita Joshi ,Computer Science, CET-IILM-AHL, Greater Noida , India.
Dr. J.S.Sodhi , Information Technology, Amity University, Noida, India.
Roopali Goel, Computer Science ,CET-IILM-AHL ,Greater Noida, India.

The basic idea behind the whole process of contextual advertising can be implement according to the above hierarchy of subprocesses According to the hierarchy of project phases the first task is Extraction of entities from the web page text and matching these entites with the ad pages and desired ad publisher will target the web pages .

III. METHODOLOGY USED

The basic idea for the methodology used in the whole process of contextual advertising can be implemented by using Extraction of the entities (Name, Organization, Place etc) According to the hierarchy of project phases the first task is Extraction of entities from the web page text .

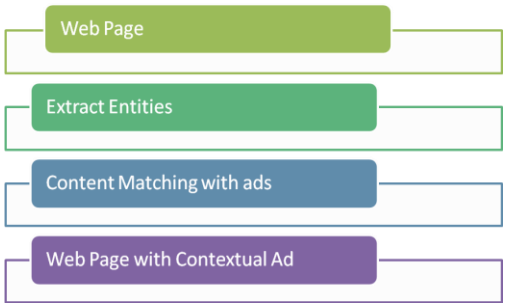

Fig 2: METHOD FOR CONTEXTUAL AD'S

PREPROCESSING (CLEANING OF TEXT)

The Preprocessing of text is required before extracting the Entities the following which are process executed sequentially over a corpus of documents:

A .Sentence analyzer and tokenizer

In order to obtain all words that are used in a given text, a tokenization process is required, i.e. Tokens are obtained by splitting a sentence along a predefined set of delimiters like spaces, commas, and dots. A token is typically a word or a digit, or punctuation. The aim is to limit the work of the tokenizer to maximize efficiency, and enable greater flexibility by placing the burden on the grammar rules, which are more adaptable. The sentence splitter is a cascade of finite-state transducers which segments the text into sentences. This module is required for the POS tagger and other modules.

B. Part of speech tagger

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. that assigns to each word a grammatical category coming from a fixed set. The set of tags includes the conventional part of speech such as noun, verb, adjective, adverb, article, conjunct, and pronoun; but is often considerably more detailed to capture many subtypes of the basic types. Parts of Speech taggers (POS taggers), also called word-category disambiguation.[1,22] ***For example:***

***Sentence*:** talks between XYZW Co. and striking Machinists union members
tokens : talks between XYZW Co. and striking Machinists union members
Pos tags: NNS IN NNP NNP CC VBG NNS NN NNS

C. Gazetteer

The terms gazetteer, lexicon and dictionary are often used interchangeably with the term list. List inclusion is a Gazetteer, or entity dictionaries, are important for performing named entity recognition (NER) accurately. The gazetteer list is the lookup list of entities. They are stored in various files which the Gazetteer uses to detect in generally the initial phases of annotations. Gazetteer prepared manually. They can categorized the different entities into different class .users can manually prepared the list of entities and it will classify different entities for entity recognition. Most studies using gazetteers for entity recognition are based on the assumption that a gazetteers is a mapping from a multiword noun to named entity categories such as

"Dr. Mehta ⟶ {Person name}"
"London ⟶ {Location}"

IV.EXTRACTION

Rule-based extraction methods are driven by hard predicates Rule-based methods are easier to interpret and develop. Therefore, rule-based systems are more useful in closed domains where human involvement is both essential and available. Many real-life extraction tasks can be conveniently handled through a collection of rules, which are either hand-coded or learnt from examples. A basic rule is of the form: "Contextual Pattern → Action".[8,9,11,22] Context pattern as the name says something related to the context. Action say's interpretation and give answer in the proper context. The action part of the rule denotes the action according to the context describe on the left hand side of the rule.For example, an extractor for contact addresses of people is created out of two phases of rule annotators: the first phase labels tokens with entity labels like people names, geographic locations like road names, city names, and email addresses. The second phase locates address blocks with the output of the first phase as additional features.

A. Rules for identifying the entity

Each rule contains set of pattern/action. The grammar always has two sides: Left and Right. The LHS of the rule contains the identified pattern that may contain regular expression. The RHS outlines the action to be taken on the detected pattern and consists of annotation manipulation statements.
LHS (regular expression for annotation pattern) ->RHS (manipulation of the annotation pattern from LHS)

B. Implementation of rule's for extraction of entity:

It defines the pattern for extraction An example of a pattern for identifying person names of the form "Dr. Roshni Mehta" consisting of a title token as listed in a dictionary of titles (containing entries like: "Prof", "Dr", "Mr"), a dot, and two capitalized words is[14,16,22]

Rule: Person Name//Matches with "Dr< name> ({Dictionary Lookup = Titles} {String = "."} {Orthography type= capitalized word}{2}) → Person Names.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Each condition within the curly braces is a condition on a token followed with an optional number indicating the repetition count of tokens.

C. Extraction of Entities

We can simplify our task by taking a example of simple text which, includes entities like person, organization, location, city, pincode.

Fig 3: Extraction of Entities

Extraction is to find words that ideally describe the content of a document. When extraction is used in information retrieval the entities should also describe the document in a way that separates it from the other documents in the collection. Entities are parts of sentences that can consist of one to several words. They are used in the same domains as extraction, but can sometimes provide more contextual information and have the ability to capture multi-word expressions. When extraction is used in information retrieval the entities should also describe the document in a way that separates it from the other documents in the collection. Extraction extracts entire sentences from a document and is used in slightly different domains. The most common use of it is in automatic summarizing of documents. Examples of such entities are:-Names of people, places, organizations, products ,E-mail addresses, URLs, Dates, numbers, sums of money, Abbreviations, Acronyms and their definition, Multiword terms.

V.CONCLUSION

In this section, we presented an overview of rule-based methods to entity extraction. We showed how rule-based systems provide a convenient method of defining extraction patterns spanning over various properties of the tokens and the context in which it resides. One key advantage of a rule-based system is that it is easy for a human being to interpret, develop, and augment the set of rules. One important component of a rule-based method is the strategy followed to resolve conflicts; many different strategies have evolved over the years but one of the most popular of these is ordering rules by priorities. Most systems allow the domain expert to choose a strategy from a set of several predefined strategies. Rules are typically hand-coded by a domain expert but many systems also support automatic learning of rules from examples. Earlier approaches are limited to some phrases which are manually entered. But using the gazetteer many words are exist for reference and recognition Through extraction of different entity can help wide area of products for ad publisher for ad marketing. With reference of multiple entities create closed group of similar entity. It helps to Present relevant data on web.

VI. FUTURE SCOPE

We present entity extraction using rule based approach. In the further approach we can the same approach for Relationship extraction. We also prepare training data which will remove stemming of data.

REFERENCES

[1] Zaiqing Nie, Ji-Rong Wen, and Wei-YingMa "Statistical Extraction from Web Manuscript" ID 0094-SIP-2011-PIEEE.R1
[2] Jasmeen Kaur, Vishal Gupta,"Effective Approaches For Extraction Of Keywords" ,IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010
[4] Massimiliano Ciaramita, Vanessa Murdock, Vassilis Plachouras,"Semantic Associations For Contextual Advertising", Journal of Electronic Commerce Research, VOL 9, NO 1, 2008
[5] Andrei Broder, Marcus Fontoura, Vanja Josifovski ,Lance Riedel" A Semantic Approach to Contextual Advertising" ,SIGIR 2007 Proceedings
[6] David Nadeau,Peter D. Turneyand Stan Matwin," Unsupervised Named- Recognition:Generating Gazetteers and Resolving Ambiguity"
[7] Stefan Dietze, Diana Maynard, Elena Demidova, Thomas Risse, Wim Peters, Katerina Doka, Yannis Stavrakas," Extraction and Consolidation for Social Web Content Preservation"
[8] Paul Bennett, Introduction to Text Categorization 20-760: Web-Based Information Architectures July 23, 2002
[9] 2004. ACE. Annotation guidelines for entity detection and tracking.
[10] Agichtein, "Extracting relations from large text collections," PhD thesis, Columbia University, 2005.
[11] E. Agichtein and V. Ganti, "Mining reference tables for automatic text segmentation," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA,2004.
[12] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plaintext
[13] Collections," in Proceedings of the 5th ACM International Conference on Digital Libraries, 2000.
[14] E. Agichtein and L. Gravano, "Querying text databases for efficient information extraction," in ICDE, 2003.
[15] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in Advances in Knowledge Discovery and Data Mining, (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds.), ch. 12, pp. 307–328, AAAI/MIT Press, 1996.
[16] J. Aitken, "Learning information extraction rules: An inductive logic programming approach," in Proceedings of the 15th European Conference on Artificial Intelligence, pp. 355–359, 2002.
[17] R. Ananthakrishna, S. chaudhuri, and V. Ganti, "Eliminating fuzzy duplicates in data warehouses," in International Conference on Very Large Databases (VLDB), 2002.
[18] R. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," Journal of Machine Learning Research, vol. 6, pp. 1817–1853, 2005.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

[18] D. E. Appelt, J. R. Hobbs, J. Bear, D. J. Israel, and M. Tyson, "Fastus: A finite-state processor for information extraction from real-world text," in IJCAI, pp. 1172–1178, 1993.
[19] A. Arasu, H. Garcia-Molina, and S. University, "Extracting structured data from web pages," in SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 337–348, 2003.
[20] S. Argamon-Engelson and I. Dagan,"Committee-based sample selection for probabilistic classifiers," Journal of Artificial Intelligence Research, vol. 11, pp. 335–360, 1999.
[21] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning,"in ICML, pp. 65–72, 2006.
[22] Sunita sarawagi "information Extractiuon" Vol. 1, No. 3 (2007) 261–377
[23] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," in FOCS '02: Proceedings of the 43rd Symposium on Foundations of Computer Science, USA, Washington, DC: IEEE Computer Society, 2002.
[24] G. Barish, Y.-S. Chen, D. DiPasquo, C. A. Knoblock, S. Minton, I. Muslea, and C. Shahabi, "Theaterloc: Using information integration technology to rapidly build virtual applications," in International Conference on Data Engineering (ICDE), pp. 681–682, 2000.
[25] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual web information extraction with lixto," in VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases, pp. 119–128, USA, San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001.

Ms. Asmita Joshi is working as a asst. prof in CSE dept. in CET-IILM-AHL Greater noida,UP,India. She is pursuing M.tech in CS From Amity University,Noida.

Dr.J.S.Sodhi Currently working as Head-IT & CIO (Assistant Vice President) with AKC Data Systems (An Amity Group and AKC Group Company).Doctorate in Information Security Management and Certified Security Compliance Specialist (CSCS).His area of interest is IT Management.

Ms. Roopali Goel is working as a Asst. prof in the department of CSE in CET-IILM-AHL Greater Noida, UP, India. She is also pursuing her M.Tech in CS(final year) from Jamia Hamdard University Delhi. Her area of interest is Cloud Computing and Wireless Networks.