Hideo Joho, Dmitry I. Ignatov (Eds.)

ECIR 2013 – Doctoral Consortium

Doctoral Consortium is associated with the 35th European Conference on Information Retrieval (ECIR 2013)
March 24, 2013, Moscow, Russia

Volume Editors

Hideo Joho
Faculty of Library, Information and Media Science
University of Tsukuba, Tsukuba, Japan

Dmitry I. Ignatov
School of Applied Mathematics and Information Science
National Research University Higher School of Economics, Moscow, Russia

Printed in National Research University Higher School of Economics.

Preface

Doctoral students were invited to the Doctoral Consortium held in conjunction with the main conference of ECIR 2013. The Doctoral Consortium aimed to provide a constructive setting for presentations and discussions of doctoral students research projects with senior researchers and other participating students. The two main goals of the Doctoral Consortium were: 1) to advise students regarding current critical issues in their research; and 2) to make students aware of the strengths and weakness of their research as viewed from different perspectives. The Doctoral Consortium was aimed for students in the middle of their thesis projects; at minimum, students ought to have formulated their research problem, theoretical framework and suggested methods, and at maximum, students ought to have just initiated data analysis.

The Doctoral Consortium took place on Sunday, March 24, 2013, at the ECIR 2013 venue, and participation is by invitation only. The format was designed as follows: The doctoral students presents summaries of their work to other participating doctoral students and the senior researchers. Each presentation was followed by a plenary discussion, and individual discussion with one senior advising researcher. The discussions in the group and with the advisors were intended to help the doctoral student to reflect on and carry on with their thesis work.

March 24, 2013

Hideo Joho
Dmitry I. Ignatov

Organization

Workshop Co-Chairs

Hideo Joho	University of Tsukuba, Tsukuba, Japan
Dmitry I. Ignatov	National Research University Higher School of Economics, Moscow, Russia

DC Mentors

Allan Hunbury	Vienna University of Technology, Austria
Dmitry I. Ignatov	Higher School of Economics, Russia
Hideo Joho	University of Tsukuba, Japan
Jaap Kamp	University of Amsterdam, The Netherlands
Natalia Loukashevitch	Moscow State University, Russia
Marie-Francine Moens	Katholieke Universiteit Leuven, Belgium
Stefan Rüger	Open University, UK
Alexander Panchenko	Université catholique de Louvain, Belgium
Konstantin Vorontsov	Computing Center for the Russian Academy of Science, Russia

Sponsoring Institutions

Russian Foundation for Basic Research
National Research University Higher School of Economics, Moscow
Yandex, Moscow

Table of Contents

Context Advertising Automation For Online Shops with a Large Commercial Base

Buzun Nazar
5th year student
nazar@ispras.ru

Institute for System Programming Russian Academy of Sciences
Moscow 109004, Russia
http://www.ispras.ru
Alytics LLC
http://alytics.ru

Abstract. The paper describes the main problems of creating and conducting advertising campaigns in systems like Yandex Direct and Google Adwords. Various their solutions are proposed. Four main steps are investigated: key phrases generation, budget forecast, banners creation and dynamic bids control. Each step is essential in such systems and requires a lot of human resources at creation or support stages. At the same time some optimisations couldn't be done manually because of huge data arrays. Here attention is focused at big online shops where automation of advertisement is particularly in demand. Proposed heuristics allow to save time and increase profit using balanced combination of automation tools, statistics collection, context agent experience and shop owner preferences.

Keywords: context automation, advertising, targeting, GSP auctions, online shops advertising

1 Motivation

The process of creating and placing advertisements in systems such as Yandex is composed of the following stages. For each product one should create a list of key phrases - phrases that would be attached to ads matched the product. Then every such phrase is added with a bid. At a time when a potential buyer enters a request into the system the ads are displayed on the following principle: each ad must contain a key phrase to be included in the search query phrase, selection for displaying and ranking of ads depends on product of bid and CTR [1] of the attached phrase. When one click on an ad the advertiser pays the price within the rules of generalized second price (GSP) auction [4].

[1] CTR is a ratio between amount clicks and amount of shows during 28 days

In manual operation via Yandex Direct or other systems to automate Yandex Direct and bids (R-broker, Elama, Seopult, yadis etc.) occur in the first place a lot of routine actions repeating from product to product during the ads creation stage. Secondly there is a need to assess the total expenditures for a given time period. Thirdly one have to dynamically update attached key phrases and their bids based on the accumulated statistical information (own and provided by an ad platform).

Most online stores use the services of systems like Yandex Market. So they can provide information about the products in the following form (YML document) which is a rather convenient format for automated generation of key phrases:

One offer per each product

```
<offer id=”111” type=”vendor.model” available=”true”>
        <url> http://www.magazin.ua/price.php?id=88521</url>
        <price>115.85</price>
        <category>Mobile</category>
        <vendor>Samson</vendor>
        <model>galaxa s</model>
        <description>...</description>
        <characteristics>...</characteristics>
</offer>
```

Using information from each offer one could form key phrases from various combinations of ”category”, ”vendor”, ”model” fields and their synonyms. In the following section the problem of key phrases formation with Yandex Direct service will be considered more formally.

2 Research questions

Research tasks are divided into four parts corresponded to key phrases generation, budget forecast, banners creation and dynamic bids control.

At the key phrases generation stage one have to discover possible synonyms of category, vendor and model for each product (offer) and then form their concatenations which could be appear in search query. Additionally obtained phrases may be extended with some specific product parameters or sale words. After that weak (non popular) phrases have to be removed. For this purpose Yandex Direct provides following helpful methods: Wordstat and GetKeywordsSuggestion which will be described in Methodology section.

To make budget forecast one need select optimal key phrases according to their CTR, pay per click, conversion, profit from corresponded product and make an assumption about the month budget required for all ads and obtained income. Here BudgetForecast method could be used to get suggestions about required phrase parameters.

Banners creation part means massive ads generation from a few amount of banner templates which contains special markers that would be replaced with an individual product information. Also here system could provide some recommendations about ads form and useful words basing on previously created ads and their popularity.

Dynamic bids control means continuous bid update to maximise profit or other function for a short period and with some time restrictions installed by Yandex Direct.

3 Methodology

The traditional way of key phrases formation in similar frameworks is applying numerous split/concatenate/translate rules for vendors, categories, models, etc. according to user choice. This process could and should be automated so as applying general rules to all products not appropriate in terms of both the quality of keywords and increasing total keywords amount.

3.1 Selection of synonyms for categories and vendors

The problem of synonyms selection for category/vendor does not have a solution with sufficient accuracy. So in this situation automation is creating the most accurate list of possible synonyms and providing a final choice to the user. Using for this purpose Yandex Direct API (Wordstat, GetKeywordsSuggestion) the task is to get maximum information about synonyms, clear excess words from returned phrases and avoid repetition (PyMorphy). The rules for the excess words filtering is determined empirically basing on their repeating. Test example for request ”hyundai” is returns following results: ”hunday”, ”huynday”, ”hundai”, ”hyndai”, ”huyndai”, ”Hyundai”, ”Hyundai”, ”xyunday”.

While GetKeywordsSuggestion gives this result: ”Hyundai, hyundai price, Hyundai prices, Hyundai dealer, dealer hyundai, dealership hyundai, Hyundaito buy, dealershipHyundai, for sale hyundai, carshyundai, for saleHyundai, authorized dealerHyundai, Hyundai, to buy hyundai, Hyundai, Hyundaisaloon, Hyundaicar, to buy Hyundai, Hyundai of St. Petersburg, Auto Show Hyundai, cars Hyundai, hyundai in St. Petersburg, Hyundai cars”

3.2 Synonym matching for product models

As an example the model of phone vendor Samson ”GT-S5230 La Fleur” is considered. Among the requests related to this product according to Wordstat statistics one may meet the following options: ”S5230 Fleur GT, GTS-star 5230, 5230 La Fleur, 5230, GTS-star, S5230 Fleur, etc”.

One way of making synonymous is partition of the original model into indivisible units (”GT”,”S”, ”5230”, ”La”,”Fleur”) and subsequent generation of all possible their combinations. It’s necessary to select only those outcomes that corresponds to sufficient number of search queries per month from potential buyers. Direct Wordstat method provide information on requests and their frequency containing a given keyword phrase that is passed as a parameter.

In case of large amount of models such method of synonyms formation could not be feasible in terms of time-consuming and limited number of calls per day established by Direct system. An alternative to this approach is smaller series of requests to Direct with the most frequent units (”GT”,”S”, ”5230”, ”La”,”Fleur”) and with possible refinement of category and vendor (”samson 5230”, ”mobile 5230 Fleur”). Using obtained queries afterwards is possible to identify some substrings in the full name of the model as an intersection with the phrases from Wordstat result. It is also necessary to identify and eliminate the requests related to other products by adding the negative keyword (in our case ”-star”).

This method significantly reduces the time but often still do not satisfies for the number of requests per day. For this purpose in current implementation products are pre-clustered (grouped by similarity) using SLPA community detection algorithm [1] which is the most effective for clustering graphs with a small cluster overlapping [2].

This approach allows to identify both common queries for multiple products in one group and make synonyms with a general rule for several models (for example if it is established that the GT-S5230 can be replaced by S5230 one could put that GT-C330 may be replaced with C330).

3.3 Budget forecast

Direct system provides following information for each key phrase for each of the three positions in the window (placement in special offers, first place in the right box, guaranteed placement in the right box): predicted CTR, minimum bid to be placed into the fixed position (p), amount of shows per month for a given region (shows). Revenue from an ad in a given position could be computed as follows:

$$shows * CTR * (\alpha * v * c - p),$$

where α is a profit from product (in percent), c is conversion.

Thus for a given total budget limit this problem is reduced to the classic ”knapsack problem” where the object is a key phrase in a fixed position. This problem could be effectively solved by ”greedy” algorithm described in [3].

In the case when α is unknown parameter the solution is divided into two stages. In the first stage the budget is distributed between product groups to maximizes the total number of ad clicks (or conversion). In the second stage within each group one is able to optimize profit (here α is constant) where the algorithm ”sequential elimination of dominated strategies” [3] is applied as more efficient than its ”greedy” analogue within a small number of objects.

3.4 GSP auctions and dynamic bids control

In accordance with bids attached to key phrases abs are placed into one of the possible positions characterized by a click probability (CTR = α_j). Considering subjective effect of an ad (γ_i) click probability of keyword i at position j is $\alpha_j \gamma_i$ Every ad and its key phrase is assigned with a score (s_i) (clicks/shows). Thus the cost of a click on an ad i in position j is assigned as follows:

$$p_j = \frac{s_{j+1} b_{j+1}}{s_i}$$

Expected profit is given by the following relation [4]:

$$\alpha_j \gamma_i (v_i - p_i) = \alpha_j \gamma_i (v_i - \frac{s_{j+1} b_{j+1}}{s_i}),$$

where v_i is expected profit from a single click.

One should take into account that frequency of search queries is much higher than bid updates. This fact occurs due to the auction manipulates with various key phrases at a time, limited computing power, incomplete information about the parameters of the auction and inability to precisely predict current scores of the other advertisers. To increase the effectiveness of advertising campaigns score values assigned with random variable $s_i = \overline{s_i} \epsilon_i$.

As a result the objective function of GSP auction participant converts to the following form [5]:

$$\underset{b_i}{max} EU_i(b_i; b_{-i}; s) v_i Q_i(b_i; b_{-i}; s) - TE_i(b_i; b_{-i}; s)$$

where Q_i is expected number of clicks per an ad, TE_i is expected advertiser consumption for clicks.

4 Progress made so far

Currently two stages key phrases generation and budget forecast are implemented. As noted above key phrases generation consists of synonyms selection for vendors, categories and models which are substrings of final key phrases.

For estimation of vendor synonyms quality ten most common vendors from the category of electronics were collected for test. The obtained precision is 86 percent and recall is 100 percent while precision value of unfiltered Yandex.Direct suggestions is less than 5 percent. The same experiment for ten most common categories resulted in recall equal to 91 percent and precision was 10 percent that nevertheless could be corrected by manual final synonyms selection. Similar experiment for product models was not conducted due to the lack of ground truth data.

As for budget forecasting proposed method (section 3.3) was compared with a standard approach which maximize total number of clicks for a fixed value of advertising costs (budget) per month. Results are presented in [Fig.1]. Our method outperforms clicks optimisation of adds in average 15 percent of final profit from all ads. Data for this experiment is partially synthetic and currently we are collecting statistics about real time advertising campaigns for a more realistic assessment.

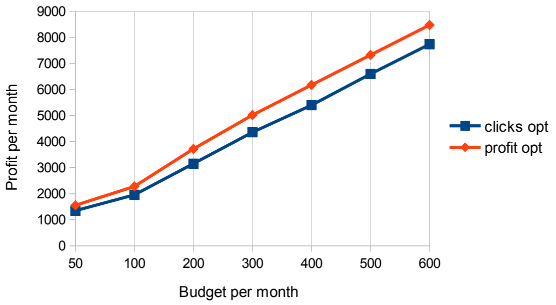

Fig. 1. Comparison of two budget forecast approaches: clicks opt is classic optimisation method that maximises amount of clicks from all keywords per month; profit opt is the method considered in this article (section 3.3)

5 Future plan

Further research consists of banners creation and dynamic bids control phases implementation. Banners creation stage constitutes a preparation of a small number of patterns which are used in huge amount of ads generation. This process could be supplemented by sale words recommendation and may be entirely automated using Markov chain based on n-gram models. Dynamic bids control is similar to budget forecasting stage but requires continuous monitoring of competitors and key phrases parameters (clicks per hour) prediction for a short time period.

References

1. J. Xie, B. Szymanski, X. Liu. 2011. SLPA: Uncovering Overlapping Communities in Social Networks via Speaker-listener Interaction Dynamic Process.
2. N. Buzun, A. Korshunov. 2012. Innovative Methods and Measures in Overlapping Community Detection.
3. N. Kuzyurin, S. Fomin. 2011. Efficient algorithms and computational complexity.
4. B. Edelman, M. Ostrovsky, M. Schwarz. 2007. Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords.
5. S. Athey, D. Nekipelov. 2010. A Structural Model of Sponsored Search Advertising Auctions.

DIADEM: Domains to Databases[*]

Andrey Kravchenko
2^{nd} year PhD student
Supervisor: Professor Georg Gottlob

Department of Computer Science, Oxford University,
Wolfson Building, Parks Road, Oxford OX1 3QD
firstname.lastname@cs.ox.ac.uk

Abstract. What if you could turn all websites of an entire domain into a single database? Imagine all real estate offers, all airline flights, or all your local restaurants' menus automatically collected from hundreds or thousands of agencies, travel agencies, or restaurants, presented as a single homogeneous dataset.
Historically, this has required tremendous effort by the data providers and whoever is collecting the data: Vertical search engines aggregate offers through specific interfaces which provide suitably structured data. The semantic web vision replaces the specific interfaces with a single one, but still requires providers to publish structured data.
Attempts to turn human-oriented HTML interfaces back into their underlying databases have largely failed due to the variability of web sources. We demonstrate that this is about to change: the availability of comprehensive entity recognition together with advances in ontology reasoning have made possible a new generation of knowledge-driven, domain-specific data extraction approaches. To that end, we introduce DIADEM, the first automated data extraction system that can turn nearly any website of a domain into structured data, working fully automatically, and present some preliminary evaluation results.
We also present a brief overview of BER_yL, DIADEM's sub-module for web block classification.

1 The DIADEM system

Most websites with offers on books, real estate, flights, or any number of other products are generated from some database. However, meant for human consumption, they make the data accessible only through, increasingly sophisticated, search and browse interfaces. Unfortunately, this poses a significant challenge in automatically processing these offers, e.g., for price comparison, market

[*] The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement DIADEM, no. 246858. Andrey is a third year DPhil student, supervised by Professor Georg Gottlob (firstname.lastname@cs.ox.ac.uk)

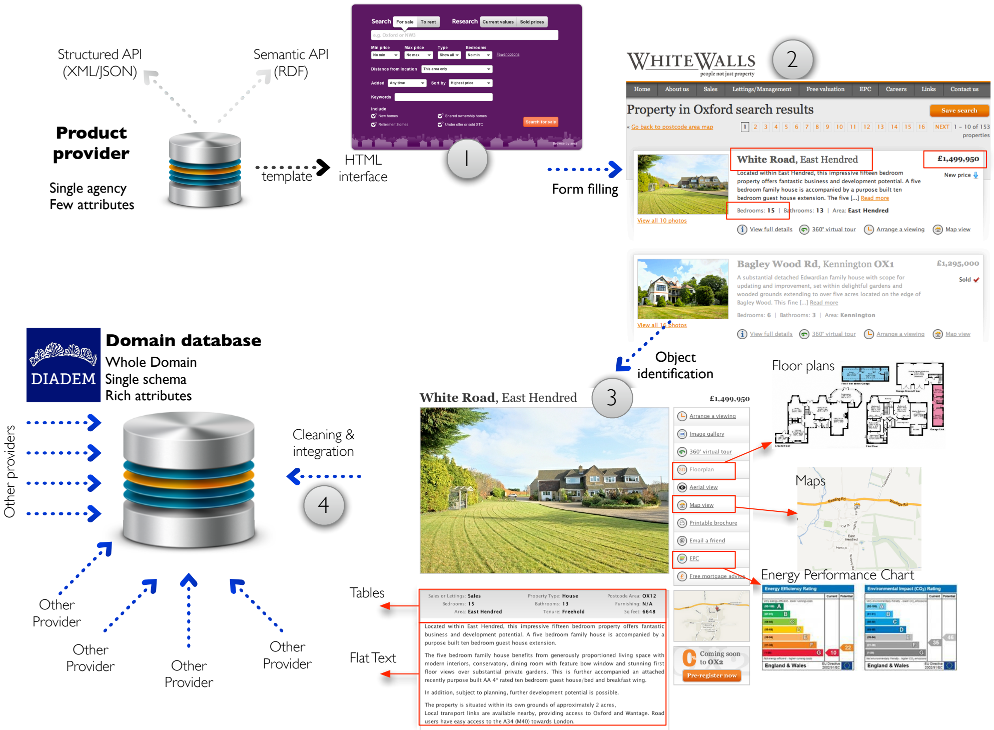

Fig. 1: Data extraction with DIADEM

analysis, or improved search interfaces. To obtain the data driving such applications, we have to explore human-oriented HTML interfaces and extract the data made accessible through them, without requiring any human involvment.

Automated data extraction has long been a dream of the web community, whether to improve search engines, to "model every object on the planet"[1], or to bootstrap the semantic web vision. Web extraction comes roughly in two shapes, namely web *information extraction* (IE), extracting facts from flat text at very large scale, and web *data extraction* (DE), extracting complex objects based on text, but also layout, page and template structure, etc. Data extraction often uses some techniques from information extraction such as entity and relationship recognition, but not vice versa. Historically, IE systems are domain-independent and web-scale [13, 14], but at a rather low recall. DE systems fall into two categories: domain-independent, low accuracy systems [15, 16] based on discovering the repeated structure of HTML templates common to a set of pages, and highly accurate, but site-specific systems [4, 5] based on machine learning.

We argue that a **new trade-off** is necessary to make *highly accurate, fully automated web extraction possible at a large scale.* We trade off scope for accuracy and automation: By limiting ourselves to a specific domain where we can provide substantial knowledge about that domain and the representation of its objects on

[1] Bing's new aim, http://tinyurl.com/77jjqz6.

web sites, automated data extraction becomes possible at high accuracy. Though not fully web-scale, one domain often covers thousands or even tens of thousands of web sites: To achieve a coverage above 80% for typical attributes in common domains, it does not suffice to extract only from large, popular web sites. Rather, we need to include objects from thousands of small, long-tail sources, as shown in [6] for a number of domains and attributes.

Figure 1 illustrates the principle of *fully automated data extraction at domain-scale*. The input is a website, typically generated by populating HTML templates from a provider's database. Unfortunately, this human-focused HTML interface is usually the only way to access this data. For instance, of the nearly 50 real estate agencies that operate in the Oxford area, not a single one provides their data in structured format. Thus data extraction systems need to explore and understand the interface designed for humans: A system needs to automatically navigate the search or browse interface **(1)**, typically forms, provided by the site to get to result pages. On the result pages **(2)**, it automatically identifies and separates the individual objects and aligns them with their attributes. The attribute alignment may then be refined on the details pages **(3)**, i.e., pages that provide comprehensive information about a single entity. This involves some of the most challenging analysis, e.g., to find and extract attribute-value pairs from tables, to enrich the information about the object from the flat text description, e.g., with relations to known points-of-interest, or to understand non-textual artefacts such as floor plans, maps, or energy performance charts. All that information is cleaned and integrated **(4)** with previously extracted information to establish a large database of all objects extracted from websites in that domain. If fed with a sufficient portion of the websites of a domain, this database provides a comprehensive picture of all objects of the domain.

That *domain knowledge* is the solution to high-accuracy data extraction at scale is not entirely new. Indeed, recently there has been a flurry of approaches focused on this idea. Specifically, domain-specific approaches use background knowledge in form of ontologies or instance databases to replace the role of the human in supervised, site-specific approaches. Domain knowledge comes in two fashions, either as instance knowledge (that "Georg" is a person and lives in the town "Oxford") or as schema or ontology knowledge (that "town" is a type of "location" and that "persons" can "live" in "locations"). Roughly, existing approaches can be distinguished by the amount of schema knowledge they use and whether instances are recognised through annotators or through redundancy. One of the dominant issues when dealing with automated annotators is that *text annotators have low accuracy*. Therefore, [7] suggests the use of a top-k search strategy on subsets of the annotations provided by the annotators. For each subset a separate wrapper is generated and ranked using, among others, schema knowledge. Other approaches exploit *content redundancy*, i.e., the fact that there is some overlapping (at least on the level of attribute values) between web sites of the same domain. This approach is used in [12] and an enumeration of possible attribute alignments (reminiscent of [7]). Also [2] exploits content redundancy, but focuses on redundancy on entity level rather than attribute level only.

Unfortunately, all of these approaches are only half-hearted: They add a bit of domain knowledge here or there, but fail to exploit it in other places. Unsurprisingly, they remain stuck at accuracies around $90 - 94\%$. There is also no single system that covers the whole data extraction process, from forms over result pages to details pages, but rather most either focus on forms, result or details pages only.

To address these shortcomings, we introduce the **DIADEM engine** which demonstrates that through domain-specific knowledge in all stages of data extraction we can indeed achieve high accuracy extraction for entire domain. Specifically, DIADEM implements the full data extraction pipeline from Figure 1 integrating form, result, and details page understanding.

The exploration phase of DIADEM is supported by the page and block classification in BER$_y$L, where we identify, next links in paginate results, navigation menus, and irrelevant data such as advertisements. We further cluster pages by structural and visual similarity to guide the exploration strategy and to avoid analysing many similar pages. Since such pages follow a common template, the analysis of one or two pages from a cluster usually suffices to generate a high confidence wrapper. We will give a detailed description of BER$_y$L in the next section.

2 The BER$_y$L sub-module

When a human looks at a web page he sees a meaningful and well-structured document, but when a computer looks at the same page the only thing it sees is HTML code and an agglomeration of rectangular boxes. Whilst it is probably infeasible for a machine to replicate the human eye directly, it would be very useful for it to understand the structure and semantics of the page for a wide range of different applications. Web-search is an especially important potential application, since structural understanding of a web page will allow to restrict link analysis to clusters of semantically coherent blocks. Hence, we aim to build a system that provides structural and semantic understanding of web pages.

In this section we are primarily concerned with the task of web block classification. Informally speaking, a web block is an area on a web page that can be identified as being visually separated from other parts of a web page and carries a certain semantic meaning, such as advertisement, navigation menu, footnote, etc. It is through the semantic meaning of individual web blocks that a human understands the overall meaning of a web page. There are a lot of blocks with a very common semantic meaning among different websites and domains (e.g. Logos, Google Maps, etc), but there are also blocks that are domain-specific, e.g. Floor Plan Images for the Real Estate domain. This diversity of blocks makes the task of their accurate and fast detection challenging from a research and implementation perspective. In general, the hardness of the block classification problem is not in the complexity of individual classifiers, but in the complexity of the entire classification system that employs individual block classifiers to enhance the accuracy and performance of other classifiers. There are several important applications to the task of web block classification, such as automatic and semi-automatic wrapper induction [26, 27], assisting the visually impaired people with going through the website's internal content [20], mobile web browsing [17, 18, 20, 22, 19, 28], ad detection, topic clustering [29], and web search [21, 23–25].

DIADEM's web block classification system is called BER$_y$L (**B**lock classification with **E**xtraction **R**ules and machine **L**earning). There are three main requirements that BER$_y$L must meet:

1. the ability to cover a diverse set of blocks, both domain-dependent (e.g. a floor plan on a result page) and domain-independent (e.g. a social interaction form);
2. acceptable precision, recall and performance rates for all blocks;
3. being easily extendable in adaptation to new block classifiers without any loss in precision and recall rates for existing classifiers and minimal loss in performance rates.

There has been a considerable amount of research done in the field of web block classification. However, most approaches attempted to classify a relatively small set of domain-independent blocks with a limited number of features, whilst we aim at classifying both domain-dependent and domain-independent blocks. Also, none of the papers we are familiar with talked about the extendability of their approaches to new block types and features. Finally, most of these approaches have precision and recall values for the majority of block types that are unacceptable for BER$_y$L, which can be partially explained by the fact that they attempt to classify different blocks with the same set of features, whist we attempt to employ individual feature sets for different types of blocks.

We thus aim at building a system that can handle block-specific features. We would also like to introduce features for new block types with sacrificing as little automation as possible. This leads us to a separation of BER$_y$L into three general components:

- classification of individual web blocks;
- large-scale classification of domain-independent web blocks;
- large-scale classification of domain-dependent web blocks.

For individial blocks we can define the features and build the training corpus in a manual way. However, outliers are very common for many block types, such as Pagination Bars, and we have to define very block-specific features to take them into account. The general problem of an entirely manual approach to web block classification is that for a lot of block types we have to generate very specific and complicated features, and some of these blocks are very diverse in their visual and structural layout, which requires a lot of data to build the training corpus, which in its turn is correlated with increased human effort. We would like to automate this process as much as possible. We hence encounter the problem of building a large scale classification system for domain-independent and domain-dependent blocks.

In the case of a large-scale classification for domain-independent blocks, such as Navigation Menus or Advertisements, the significant number of blocks, many of which come in a diverse structural and visual representation, makes the set of potential features so large that the feature extraction process can become a bottleneck for the system. This will leave us to either contracting the feature space or optimising the performance of the feature extraction process. We would also like to employ techniques for minimising the training data, such as semi-supervised learning, as the number of labelled instances required for the classifier to work can be very large, even given a diverse and representative set of features.

In the case of a large-scale classification for domain-dependent blocks, such as Floor Plans for the Real Estate domain or Commentary Sections for the Blogs and Online Forums domain, the space of potential features becomes so large, if not infinite, that we will have to resort to techniques for the automatic learning of new features, such as Genetic Algorithms.

For both domain-dependent and domain-independent blocks we would like to combine human generated rules (e.g. "Floor Plans can only appear on the Real Estate result pages") and the training data. If the training data strongly contradict the rule, we can then remove that rule from BER$_y$L.

The task of web block classification is technically challenging due to the diversity of blocks in terms of their internal structure (representation in the DOM tree and the visual layout) and the split between domain-dependent and domain-independent blocks. Hence, from a technical perspective, it is important to have a **global feature repository** that can provide a framework for defining new features and block type classifiers through a template language. The user of BER$_y$L will be able to extend it with new features and classifiers easily by generating them from existing templates, rather than writing them from scratch. That will make the whole hierarchy of features and classifiers leaner, as well as making the process of defining new block types and their respective classifiers more straightforward and less time-consuming. Ideally, we would like to generate new block classifiers in a fully automated way, such that given a set of structurally and visually distinct web blocks of the same type, the block classification system would be able automatically to identify the optimal list of features to describe that block, taking some of those from the existing repository and generating the new ones that have not existed in the repository beforehand. However, it would be almost infeasible to go with this approach in the case of BER$_y$L, since the diversity of web blocks we want to classify is likely to cause the space of potential features to be extraordinarily large if not infinite. Hence we will have to limit ourselves to a semi-automated approach to the generation of new features and block type classifiers.

The contributions of the approach we take in BER$_y$L are two-fold, namely, improving the quality of classification and minimising the effort of generating new classifiers (as described in the paragraph above). Contributions 1-3 and 4-6 refer to the quality of classification and generation aspects respectively.

1. We provide a holistic view of the page that gives a coherent perspective of the way it is split into web blocks and the way in which these blocks interact.

2. We employ domain-specific knowledge to enhance performance of both domain-dependent (e.g. Floor Plans for the Real Estate domain) and domain-independent (e.g. Navigation Menus) classifiers.
3. We provide a global feature repository that allows the user of BER$_y$L to add new features and block classifiers easily.
4. The holistic view of the page is implemented through a system of mutual dependencies and constraints. We also run the classifiers in an organised schedule, which is selected in such a way that it maximises the accuracies of the individual classifiers as well as the overall performance of BER$_y$L.
5. We encode domain-specific knowledge through a set of logical rules, e.g. that for the Real Estate and Used Cars domains the Navigation Menu is always at the top or bottom, and rarely to the side of the main content area of the page, or that for the Real Estate domain Floor Plans can only be found within the main content area of the page.
6. The global feature repository is implemented through baseline global features and a logic-based template language used to derive local block-specific features. We also use textual annotations from the global annotation repository as features for web block classification to enhance the accuracy of individual classifiers further.

Web block classification It does not seem feasible to solve the block classification problem through a set of logic-based rules, as it is often the case that there are a lot of potential features that can be used to characterise a specific block type, but only a few play a major role in uniquely distinguishing it from all other block types. Some of these features are continuous (e.g. the block's width and height), and it can be hard for a human to specify accurate threschold boundaries manually. Hence, in BER$_y$L we decided to use a machine learning (ML) approach for web block classification.

References

1. Benedikt, M., Gottlob, G., Senellart, P.: Determining relevance of accesses at runtime. In: *PODS.* (2011)
2. Blanco, L., Bronzi, M., Crescenzi, V., Merialdo, P., Papotti, P.: Exploiting Information Redundancy to Wring Out Structured Data from the Web. In: *WWW.* (2010)
3. Crescenzi, V., Mecca, G.: Automatic information extraction from large websites. *J. ACM* **51**(5) (2004) 731–779
4. Zheng, S., Song, R., Wen, J.R., Giles, C.L.: Efficient record-level wrapper induction. In: *CIKM.* (2009)
5. Dalvi, N., Bohannon, P., Sha, F.: Robust web extraction: an approach based on a probabilistic tree-edit model. In: *SIGMOD.* (2009)
6. Dalvi, N., Machanavajjhala, A., Pang, B.: An analysis of structured data on the web. In: *VLDB.* (2012)
7. Dalvi, N.N., Kumar, R., Soliman, M.A.: Automatic wrappers for large scale web extraction. In: *VLDB.* (2011)

8. Dragut, E.C., Kabisch, T., Yu, C., Leser, U.: A hierarchical approach to model web query interfaces for web source integration. In: *VLDB.* (2009)
9. Furche, T., Gottlob, G., Grasso, G., Guo, X., Orsi, G., Schallhart, C.: Opal: automated form understanding for the deep web. In: *WWW.* (2012)
10. Furche, T., Gottlob, G., Grasso, G., Orsi, G., Schallhart, C., Wang, C.: Little knowledge rules the web: Domain-centric result page extraction. In: *RR.* (2011)
11. Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.: Oxpath: A language for scalable, memory-efficient data extraction from web applications. In: *VLDB.* (2011)
12. Gulhane, P., Rastogi, R., Sengamedu, S.H., Tengli, A.: Exploiting content redundancy for web information extraction. In: *VLDB.* (2010)
13. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: open information extraction on the web. In: *NAACL.* (2007)
14. Lin, T., Etzioni, O., Fogarty, J.: Identifying interesting assertions from the web. In: *CIKM.* (2009)
15. Simon, K., Lausen, G.: Viper: augmenting automatic information extraction with visual perceptions. In: *CIKM.* (2005)
16. Liu, W., Meng, X., Meng, W.: Vide: A vision-based approach for deep web data extraction. *TKDE.* **22** (2010) 447–460
17. Romero, S., Berger, A.: Automatic partitioning of web pages using clustering. In: *MHCI.* (2004)
18. Xiang, P., Yang, X., Shi, Y.: Effective page segmentation combining pattern analysis and visual separators for browsing on small screens. In: *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.* (2006)
19. Baluja, S.: Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework. In: *WWW'06 Proceedings of the 15th international conference on World Wide Web, 33–42.* (2006).
20. Gupta, S., Kaiser, G., Neistadt, D., Grimm, P.: DOM-based content extraction of HTML documents. In: *Proceedings of the twelfth international conference on World Wide Web WWW'03.* (2003).
21. Wu, C., Zeng, G., Xu, G.: A web page segmentation algorithm for extracting product information. In: *Proceedings of the 2006 IEEE International Conference on Information Acquisition.* (2006).
22. Xiang, P., Yang, X., Shi, Y.: Web page segmentation based on Gestalt theory. In: *2007 IEEE International Conference on Multimedia and Expo.* (2007).
23. Yu, S., Cai, D., Wen, J., Ma, W.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: *WWW'03.* (2003).
24. Cai, D., Yu, S., Wen, J., Ma, W.: Block-based web search. In: *SIGIR'04.* (2004).
25. Cai, D., He, X., Wen, J., Ma, W.: Block-level link analysis. In: *SIGIR'04.* (2004).
26. Zheng, S., Song, R., Wen, J., Giles C.L.: Efficient record-level wrapper induction. In: *CIKM'09*, (2009).
27. Wang, J., Chen, C., Wang, C., Pei, J., Bu, J., Guan, Z., Zhang, W.V.: Can we learn a template-independent wrapper for news article extraction from a single traning site? In: *KDD'09*, (2009).
28. Maekawa, T., Hara, T., Nishio, S.: Image classification for mobile web browsing. In: *WWW 2006*, (2006).
29. Luo, P., Lin, F., Xiong, Y., Zhao, Y., Shi, Z.. Towards combining web classification and web information extraction: a case study. In: *KDD '09*, (2009).

Semantic Role Labeling System for Russian Language

Ilya Kuznetsov
1st year PhD
Higher School of Economics, Faculty of Philology
Supervisor: Anastasia Bonch-Osmolovskaya

iokuznetsov@gmail.com

Abstract. Semantic Role Labeling in narrow sense can be formulated as assigning semantically motivated labels from some inventory to the arguments of a predicate. We develop an architecture and methodology for building a machine learning-based SRL system for a resource-poor language incl. unsupervised semantics analysis and subcategorization frame induction based on Russian material.

Keywords: Semantic role labeling, semantic role induction, lexical similarity modeling

1 Motivation

Semantic Role Labeling (SRL) is one of the most popular natural language processing tasks today. SRL in narrow sense can be formulated as assigning semantically motivated labels (or roles) from some inventory to the arguments of a predicate based on the information about predicates' semantics and usage. Solving the SRL task makes it possible to build object-predicate-based representations of raw texts, which is enough for solving most of practical unstructured text mining tasks.

Building a successful SRL system requires a lot of additional linguistic resources such as syntactic parsers, thesauri, large usage corpora etc. That makes it hard for resource-poor languages like Russian to achieve the state-of-art performance obtained for English. Creating such linguistic resources by hand requires a lot of time and effort. One way to speed up the development of SRL for resource-poor languages is to use methods for automatically obtaining the data required. The aim of the proposed research is to solve a variety of SRL-related tasks (along with the narrow SRL task itself) and to obtain a full-featured SRL system for Russian language as result.

From the methodological point of view the research will be helpful for developing the SRL systems for other resource-poor languages. The results can be also applied to solve various cross-language data mining and machine translation tasks.

2 Overview

A minimalistic architecture for a semantic role labeling system was presented in the seminal work of D. Gildea and D. Jurafsky ([1]). The SRL task itself was formulated as a classification problem which could be solved with machine learning methods. Given the input sentence, target predicate word, it's argument constituents and relevant role set, the system should be able to assign each constituent a role, which could be abstract (like "*Agent*" or "*Theme*") or specific (like "*Speaker*" or "*Buyer*") depending on the selected description framework and corresponding role sets.[1] We will stick to that architecture and task definition while building a basic SRL system, which can be then extended by additional pre- and postprocessing modules like Named Entity processor ([2]), entity- and event-level coreference resolution ([3]) or a rule-based semantic inference engine ([4]).

The majority of lexical and corpus resources used as input in Gildea and Jurafsky's basic SRL algorithm do not have good elaborated analogues in Russian, so we have to replace them with automatically obtained ones. This primarily concerns a database like FrameNet [5] or PropBank[6], which would describe predicates in terms of their role sets and contain large amounts of usage data needed to train a classifier. A Russian FrameBank [7] is developed and could provide some primary test and training data, which could be then used to bootstrap a bigger corpus.

Another type of resources Russian lacks is thesauri. As it was shown in [1], the most accurate feature combination for argument classification is obtained by just using the lemma of argument combined with the predicate's lemma. Let's suppose we use a bag-of-words representation of a sentence (like [*buy*, *Peter*, *1000$*, *an apple*]) and operate just with lexical meanings. With this information only, we're often able to correctly assign the roles. However, that feature has very low coverage due to the data sparseness, that is, it's likely that the predicate-noun combination we meet in an input sentence isn't presented in training data. To cope with that low coverage issue, an external knowledge about semantic relatedness between words can be used. Two main approaches for generalizing over the lexemes are using a WordNet [8] and performing clustering of lexicon on a large amount of data. The former method provides more accurate information about word relatedness, while the latter has better coverage and can be tuned for using in a specific domain. The experiments show, that even on English material using clusters does not decrease performance compared to WordNet ([1]), which shows that it is possible to build a SRL system even without hand-elaborated static resources.

Considering the problems described above, the architecture of our Semantic Role Labeling system is organized as shown in Fig. 1. Given an input raw sentence and a description of the verb under concern, the modules of a system perform the following steps.

- Preprocessing
  - Tokenization

[1] See [1] and [9] for a detailed overview

  - Sentence boundaries detection
  - Morphological analysis and lemmatization
  - Syntactical parsing
- Finding relevant arguments in a dependency tree based on semi-automatically collected information about predicate usage patterns
- Fetching the lexical information about arguments using automatically built lexical similarity clusters (or just a similarity matrix)
- Classifying the arguments with regard to the input role set using the information about role coding from a large corpus.

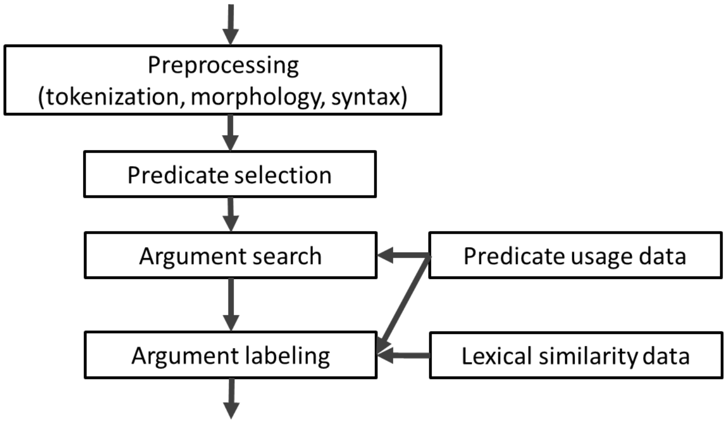

Fig. 1. Basic SRL system architecture

The next section describes methodological aspects of the last three steps, while the preprocessing is performed by external resources and integrating these in the SRL system seems to be a purely technical problem.

3 Planned Methodology

3.1 Measuring the Feature Impact

It seems rational to first determine which features are most salient for semantic role labeling. We plan to follow the guidelines given in [1] and measure the probability of assigning a correct label to a syntax tree node based on various combinations of input features. The gold standard data for this experiment will be taken from the Russian FrameBank. Gildea and Jurafsky show that the most salient SRL features for English are target word (predicate itself), voice (active or passive), argument's head word, argument's phrase type and governing category, relative position (left or right) and the path in syntactical tree. This result should be proved for Russian, because of the typological differences between these languages. For example, free word order in

Russian should significantly weaken the position feature; on the contrary, the morphological feature should become stronger because in Russian a lot of syntactical information is encoded in morphology. In addition, we plan to use a labeled dependency tree representation, which is more adequate for Russian syntax than the constituent model. We should adapt standard feature sets and re-evaluate them on Russian material to understand which external components require detailed elaboration.

3.2 Finding the Arguments

Finding the arguments is usually formulated as a classification problem, where for each sentence constituent it is decided, whether or not it belongs to a predicate's argument set. In dependency tree paradigm the relation between a predicate and its arguments is marked up by a parser, so the problem can be stated as separating obligatory arguments from optional ones. The proposed method consists of following steps:

- Inducing the subcategorization frames for predicates based on a large automatically parsed corpus (while the data from Russian FrameBank is insufficient). As result we build a model which classifies the input sentence arguments based on the information about a predicate's usage patterns.
- Evaluating the subcategorization frames using the Russian FrameBank as gold standard and finding the best algorithm for argument frame detection.
- Clustering the predicates based on their subcategorization frames (which include syntactical and morphological information) and lexical argument fillers
- Using the information about general usage patterns when dealing with an unknown predicate.

The overall strategy at this step is to maximize recall in order to deliver all core arguments to the Semantic Role Labeling stage.

3.3 Lexical Information about Arguments

Although the impact of lexical features on semantic role assignment should be evaluated, it seems natural, that these features will play an important role in classification process. Lexical similarity can be modeled by a WordNet-like resource, a clustering algorithm (e.g. [10]) or a more flexible distributional similarity representation, for example, a simple similarity matrix based on predicate-argument distribution over a large corpus could be used. The latter seems to be promising in context of semantic role labeling. It is not obvious that the aspect of similarity captured by ontological, hierarchical or other static group-based representation is the one that would be of use for labeling the arguments with semantic roles. The individual compatibility of predicates and their argument lexemes could be modeled with a large set of binary features like *[+break]*, *[+crack]* assigned to lexemes. Based on that information we could induce predicate distribution-based relationships like *[+break]* & *[+crack]* → *[+smash]* (that is, if a lexeme collocates as direct object with predicates *break* and *crack*, then it’s likely to combine with a predicate *smash* in the direct object position). However, that hypothesis should be tested and refined during an experiment which consists of the following:

- Building binary distributional similarity matrices of argument lexemes using different feature sets (e.g. *verb + relation_label*, *verb + morphology*, *collocating adjectives* etc.).
- Performing clustering using different feature sets
- Obtaining a well-described set form Russian WordNet
- Measuring the degree of correspondence of lexemes marked as filling the same role by the same predicate taken from Russian FrameBank based on each method described above and selecting the best-matching one.

3.4 Semantic Role Labeling

The Semantic Role Labeling itself is seen now as a long-run prospect. The overall methodology is described above in the Architecture overview section, the specific set of features that would perform best on Russian material will be determined during one of the experiments. The algorithm will take dependency tree argument nodes, target predicate, lexical and morphological information as input and will perform multiclass classification over a set of roles relevant to the predicate. The examples from Russian FrameBank corpus will be used as training data. Role generalization will be performed to cope with the relatively small dataset size. It is also possible to bootstrap additional data if the corpus itself will lead to poor results due to small dataset size.

4 Progress and Future Plans

The following is already done or in progress:

- Overall methodology and architecture has been developed
- Technical platform including tokenizer, morphological analyzer, syntactic parser and a database-driven corpus access engine has been established
- Feature impact measurement experiment in progress
- Lexical similarity modeling experiment in progress

The future plans consist of:

- Refining the methodology based on feature ranking obtained during the feature impact measurement experiment
- Developing an argument search algorithm based on methodology described above
- Developing an SRL feature set and algorithm
- Component integration, overall system construction

The results obtained during the experiments will be used for speeding up the development of corresponding static lexical and corpus resources.

References

1. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics, 28(3). 2002.
2. Surdeanu, M., Harabagiu, S., Williams J., Aarseth, P.: Using predicate-argument structures for information extraction. Proceedings of ACL 2003, vol.1. 2003.
3. Lee, H., Recasens, M., Chang, A.X., Surdeanu, M., Jurafsky, D.: Joint Entity and Event Coreference Resolution across Documents. EMNLP-CoNLL. 2012.
4. "Surdeanu, M., McClosky, D., Smith, M.R., Gusev, A., Manning, C.D.: Customizing an information extraction system to a new domain. ACL HLT 2011, 2. 2011."
5. Baker, C. F.,Fillmore, C. J., Lowe, J.B.: The Berkeley FrameNet Project. Proceedings of COLING-ACL 98. 1998.
6. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics, 31(1). 2005.
7. Lyashevskaya, O.: Bank of Russian constructions and valencies. LREC 2010. 2010.
8. Fellbaum C.(ed.): WordNet. An electronic lexical database. Cambridge, MA: MIT Press. 1998.
9. Xue, N., Palmer, M.: Automatic semantic role labeling for Chinese verbs. Proceedings of the 19th International Joint Conference on Artificial Intelligence. 2005.
10. Lin, D.: Automatic Retrieval and Clustering of Similar Words. COLING-ACL98. 1998.

Semantic Models for Answer Re-ranking in Question Answering

Piero Molino

2nd year PhD. student
Dept. of Computer Science - University of Bari Aldo Moro
Via Orabona, I-70125, Bari, Italy
Advisor: Pasquale Lops
piero.molino@uniba.it

Abstract. This paper describes a research aimed at unveiling the role of Semantic Models into Question Answering systems. The objective is to use Semantic Models for answer re-ranking in order to improve the passage retrieval performance and the overall performance of the system. Semantic Models use concepts rather than simple words to represent texts, expressing them in explicit or implicit ways. This allows to compute relatedness between users' questions and candidate answers to provide better answer re-ranking. This is done by exploiting different properties, like explicit relations between concepts or latent similarities between words expressed as the similarity of the contexts in which they appear. We want to find out if the combination of different semantic relatedness measures by means of Learning to Rank algorithms will show a significant improvement over the state-of-the-art. We carried out an initial evaluation of a subset of the semantic models on the CLEF2010 QA dataset, proving their effectiveness.

1 Introduction

The task of Question Answering (QA) is to find correct answers to users' questions expressed in natural language. Much of the work in QA has been done on factoid questions, where answers are short excerpts of text, usually named entities, dates or quantities. In the last few years non-factoid QA received more attention. It focuses on causation, manner and reason questions, where the expected answer has the form of a passage of text. The passage retrieval step is, anyway, fundamental in both factoid and non-factoid QA as in the former the answers are extracted from the obtained passages, while in the latter the passage corresponds to the candidate answer itself, even if the length of the passage for non-factoid QA is much larger as shown in [21].

The presence of annotated corpora from Text REtrieval Conference (TREC) and Cross Language Evaluation Forum (CLEF) allows to use machine learning techniques to tackle the problem of ranking the passages for further extraction in factoid QA [1]. In non-factoid QA the training data adopted are of different types, like hand annotated answers from Wikipedia [22], small hand built corpora [7], Frequently Asked Questions lists [2, 18] and Yahoo! Answers Extracted corpus [19]. This allows the adoption of Learning to Rank (MLR) algorithms in order to output a sensible ranking of the candidate answers. In [21] the adoption of linguistically motivated features is shown to be effective for the task, while in [23] different MLR algorithms were compared over the same set of features. The importance of semantic features, in the form of semantic role labelling features, was shown in [3], while a comprehensive large scale evaluation, alongside with the introduction of new features, was carried out in [19].

There are still different possible semantic features that have not been taken into account. For example features coming from Distributional Semantic Models (DSMs) [20], Explicit Semantic Analysis (ESA) [6], Latent Dirichlet Allocation (LDA) [4] induced topics have never been applied to the task.

The questions this research wants to answer are:

- Do semantic features bring information that is not present in the bag-of-words and syntactic features?
- Do they bring different information or does it overlap with that of other features?
- Are additional semantic features useful for answer re-ranking? Does their adoption improve systems' performances?
- Which of them is more effective and under which circumstances?
- Is there any MLR algorithm that exploits semantic features more than others (has more relative or absolute improvement by their adoption) and why?

We think that SM can have a significant role in improving current state-of-the-art systems' performances.

2 Methodology

We are going to test if these insights are correct starting from the design and implementation of a QA framework that helps us to set up several systems with different settings. We already built the cornerstone: QuestionCube is a multilingual QA framework created using Natural Language Processing and Information Retrieval techniques. Question analysis is carried out by a full-featured NLP pipeline. The passage search step is carried out by Lucene, a standard off-the-shelf retrieval framework that allows TF-IDF, Language Modeling and BM25 weighting. The question re-ranking component is designed as a pipeline of different scoring criteria. We derive a global re-ranking function combining the scores with CombSum. More details on the framework and a description of the main scorers is reported in [11]. The next step is the implementation of different MLR algorithms in order to combine the features obtained by the scoring criteria with linear and non-linear models.

As a proof of concept we implemented some scoring criteria based on DSMs in order to realize if their adoption as alone rankers or combined with simple similarity and density criteria would improve ranking over the one obtained with classic Information Retrieval weighting schemes.

Distributional Semantic Models (DSMs) represent word meanings through linguistic contexts. The meaning of a word can be inferred by the linguistic contexts in which the word occurs. The idea behind DSMs can be summarized as follows: if two words share the same linguistic context they are somehow similar in meaning. For example, in analyzing the sentences “drink a glass of wine” and “drink a glass of beer”, we can assume that the words “wine” and “beer” have a similar meaning. Using that assumption, the meaning of a word can be expressed by the geometrical representation in a *semantic space*. In this space a word is represented by a vector whose dimensions correspond to linguistic contexts surrounding the word. The word vector is built analyzing (e.g. counting) the contexts in which the term occurs across a corpus. Some definitions of context may be the set of co-occurring words in a document, in a sentence or in a window of surrounding terms. The earliest and simplest formulation of such a space stems from the use of the Vector Space Model in IR [14]. Semantic space scalability and independence from external resources resulted in their practical use in many different tasks. For example they have been applied in several linguistic and cognitive tasks, such as synonyms choice [10], semantic priming [8,10], automatic construction of thesauri [16] and word sense induction [15].

Our DSMs are constructed over a co-occurrence matrix. The linguistic context taken into account is a window w of co-occurring terms. Given a reference corpus, the collection of documents indexed by the QA system, and its vocabulary V, a $n \times n$ co-occurrence matrix is defined as the matrix $\mathbf{M} = (m_{ij})$ whose coefficients $m_{ij} \in \mathbb{R}$ are the number of co-occurrences of the words t_i and t_j within a predetermined distance w. The $term \times term$ matrix $\mathbf{M}$, based on simple word co-occurrences, represents the simplest semantic space, called Term-Term co-occurrence Matrix (TTM). In literature, several methods to approximate the original matrix by rank reduction have been proposed. The aim of these methods varies from discovering high-order relations between entries to improving efficiency by reducing its noise and dimensionality. We exploit three methods for building our semantic spaces: Latent Semantic Analysis (LSA) [5], Random Indexing (RI) [9] and LSA over RI ($LSARI$) [17]. $LSARI$ applies the SVD factorization to the reduced approximation of $\mathbf{M}$ obtained through RI. All these methods produce a new matrix $\hat{\mathbf{M}}$, which is a $n \times k$ approximation of the co-occurrence matrix $\mathbf{M}$ with n row vectors corresponding to vocabulary terms, while k is the number of reduced dimensions. More details can be found in [12].

We integrated the DSMs into the framework creating a new scorer, the **Distributional Scorer**, that represents both question and passage by applying the addition operator to the vector representation of terms they are composed of. Furthermore, it is possible to compute the similarity between question and passage by exploiting the cosine similarity between vectors using the different matrices. The simple scorers employed alongside with the ones based on DSMs in the evaluation are: **Terms Scorer**, **Exact Sequence Scorer** and **Density Scorer**, a scorer that assigns a score to a passage based on the distance of the question terms inside it. All the scorers have an enhanced version which adopts the combination of lemmas and PoS tags as features instead of simple words.

3 Evaluation

The goal of the evaluation is twofold: (1) proving the effectiveness of DSMs into our question answering system and (2) providing a comparison between the several DSMs.

The evaluation has been performed on the *ResPubliQA 2010 Dataset* adopted in the *2010 CLEF QA Competition* [13]. The dataset contains about 10,700 documents of the European Union legislation and European Parliament transcriptions, aligned in several languages including English and Italian, with 200 questions. The adopted metric is the accuracy $a@n$ (also called $success@n$), calculated considering only the first n answers. If the correct answer occurs in the top n retrieved answers, the question is marked as correctly answered. In particular, we take into account several values of n =1, 5, 10 and 30. Moreover, we adopt the Mean Reciprocal Rank (MRR) as well, that considers the rank of the correct answer. The framework setup used for the evaluation adopts Lucene as document searcher, and uses a NLP Pipeline made of a stemmer, a lemmatizer, a PoS tagger and a named entity recognizer. The different DSMs and the classic TTM have been used as scorers alone, which means no other scorers are adopted, and combined with a standard scorer pipeline. The composition of the standard pipeline includes the Simple Terms (ST), the Enhanced Terms (ET), the Enhanced Density (ED) and the Exact Sequence (E) scores. Moreover, we empirically chose the parameters for the DSMs: the window w of terms considered for computing the co-occurrence matrix is 4, while the number of reduced dimensions considered in LSA, RI and LSARI is equal to 1000.

The performance of the standard pipeline, without the distributional scorer, is shown as a baseline. The experiments have been carried out both for English and Italian. Results are shown in Table 1, witch reports the accuracy $a@n$ computed considering a different number of answers, the MRR and the significance of the results with respect to both the baseline (†) and the distributional model based on TTM (‡). The significance is computed using the non-parametric Randomization test. The best results are reported in bold.

Considering each distributional scorer on its own, the results prove that all the proposed DSMs are better than the TTM, and the improvement is always significant. The best improvement for the MRR in English is obtained by LSA (+180%), while in Italian by LSARI (+161%). As for the distributional scorers combined with the standard scorer pipeline, the results prove that all the combinations are able to overcome the baseline. For English we obtain an improvement in MRR of about 16% compared to the baseline and the result obtained by the TTM is significant. For Italian, we achieve a even higher improvement in MRR of 26% compared to the baseline using LSARI. The slight difference in performance between LSA and LSARI proves that LSA applied to the matrix obtained by RI produces the same result of LSA applied to TTM, but requiring less computation time, as the matrix obtained by RI contains less dimensions than the TTM matrix. Finally, the improvement obtained considering each distributional scorers on its own is higher than their combination with the standard scorer pipeline.

Table 1. Evaluation Results for both English and Italian

		English					Italian				
	Run	a@1	a@5	a@10	a@30	MRR	a@1	a@5	a@10	a@30	MRR
alone	TTM	0.060	0.145	0.215	0.345	0.107	0.060	0.140	0.175	0.280	0.097
	RI	0.180	0.370	0.425	0.535	0.267‡	0.175	0.305	0.385	0.465	0.241‡
	LSA	**0.205**	**0.415**	**0.490**	0.600	**0.300**‡	0.155	0.315	0.390	0.480	0.229‡
	LSARI	0.190	0.405	**0.490**	**0.620**	0.295‡	**0.180**	**0.335**	**0.400**	**0.500**	**0.254**‡
combined	*baseline*	*0.445*	*0.635*	*0.690*	*0.780*	*0.549*	*0.445*	*0.635*	*0.690*	*0.780*	*0.549*
	TTM	0.535	0.715	0.775	0.810	0.614	0.405	0.565	0.645	0.740	0.539†
	RI	0.550	0.730	0.785	**0.870**	**0.637**†‡	0.465	**0.645**	**0.720**	**0.785**	0.555†
	LSA	**0.560**	0.725	**0.790**	0.855	**0.637**†	0.470	**0.645**	0.690	**0.785**	0.551†
	LSARI	0.555	**0.730**	**0.790**	**0.870**	0.634†	**0.480**	0.635	0.690	**0.785**	**0.557**†‡

4 Future Plan

There are several future steps to follow in order to answer the research questions. We discovered that some of the semantic features we want to adopt can be useful, but we still don't know how effective they can be inside a MLR setting, so the first improvement to make is to add the MLR algorithms for re-ranking after having gathered the features from the different scorers. This will also probably lead to even better performances than the achieved ones.

The following step will be to experiment further the usefulness of semantic features adding more of them, exploiting ESA, LDA and even more semantic models, and also incorporating other state-of-the-art linguistic features. Other vector operations for combining vectors coming from the applied DSMs will be investigated.

Once all the features are there, MLR algorithm comparison will be carried out, in order to find out which algorithms take more advantage from the semantic features. An ablation test will be useful to understand how much of the improvement is obtained thanks to the semantic features.

Alongside with those steps, different datasets will be collected, focusing mail on non-factoid QA. The reason for that is the possibility to compare our evaluation directly to the state-of-the-art ones in order to realize if the semantic features can lead to better results also on those datasets.

References

1. Agarwal, A., Raghavan, H., Subbian, K., Melville, P., Lawrence, R.D., Gondek, D., Fan, J.: Learning to rank for robust question answering. In: CIKM. pp. 833–842 (2012)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on

Web Search and Data Mining. pp. 183–194. WSDM '08, ACM, New York, NY, USA (2008)

3. Bilotti, M.W., Elsas, J.L., Carbonell, J.G., Nyberg, E.: Rank learning for factoid question answering with linguistic and semantic constraints. In: CIKM. pp. 459–468 (2010)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
6. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: In Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611 (2007)
7. Higashinaka, R., Isozaki, H.: Corpus-based question answering for why-questions. In: In Proceedings of IJCNLP. pp. 418–425 (2008)
8. Jones, M.N., Mewhort, D.J.K.: Representing word meaning and order information in a composite holographic lexicon. Psychological Review 114(1), 1–37 (2007)
9. Kanerva, P.: Sparse Distributed Memory. MIT Press (1988)
10. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240 (1997)
11. Molino, P., Basile, P.: QuestionCube: a Framework for Question Answering. In: Amati, G., Carpineto, C., Semeraro, G. (eds.) IIR. CEUR Workshop Proceedings, vol. 835, pp. 167–178. CEUR-WS.org (2012)
12. Molino, P., Basile, P., Caputo, A., Lops, P., Semeraro, G.: Exploiting distributional semantic models in question answering. In: ICSC. pp. 146–153 (2012)
13. Penas, A., Forner, P., Rodrigo, A., Sutcliffe, R.F.E., Forascu, C., Mota, C.: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In: Braschler, M., Harman, D., Pianta, E. (eds.) Working notes of ResPubliQA 2010 Lab at CLEF 2010 (2010)
14. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18, 613–620 (November 1975)
15. Schütze, H.: Automatic word sense discrimination. Comput. Linguist. 24, 97–123 (March 1998), http://portal.acm.org/citation.cfm?id=972719.972724
16. Schütze, H., Pedersen, J.O.: Information retrieval based on word senses. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175 (1995)
17. Sellberg, L., Jönsson, A.: Using random indexing to improve singular value decomposition for latent semantic analysis. In: LREC (2008)
18. Soricut, R., Brill, E.: Automatic question answering using the web: Beyond the factoid. Inf. Retr. 9(2), 191–206 (Mar 2006)
19. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers to non-factoid questions from web collections. Computational Linguistics 37(2), 351–383 (2011)
20. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res. (JAIR) 37, 141–188 (2010)
21. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.A.: Using syntactic information for improving why-question answering. In: COLING. pp. 953–960 (2008)
22. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.A.: What is not in the bag of words for *why*-qa? Computational Linguistics 36(2), 229–245 (2010)

23. Verberne, S., van Halteren, H., Theijssen, D., Raaijmakers, S., Boves, L.: Learning to rank for *why*-question answering. Inf. Retr. 14(2), 107–132 (2011)

Collecting and Analyzing Data in order to Develop and Implement Municipal Programmes and Policies

Molyarenko Olga
1st year PhD
National Research University Higher School of Economics, Moscow
Kordonsky Simon (HSE tenured professor, Department of Local Administration Head)

omolyarenko@hse.ru

Abstract. The increasing attention in modern public administration is paid to the transparency. However, it is supposed that there is an imbalance between improving the mechanisms of opening the information about public authorities' activities and carried out by them collecting socio-economic data. The tools of gaining information about the current situation, needs and urgent issues seem to lag behind significantly, and the consequences of using unauthentic data as a foundation for managerial decision are obvious. Therefore the research is dedicated to description and evaluation the current methods and techniques of collecting and analyzing data by local authorities and developing the model of relevant "intelligence". The aggregate model would be based on the current official statistics system and analysis of information space (presented in the Internet) combination.

1 Motivation

During the studying at the University (the faculty of Public Administration) we have considered a lot of unsuccessful examples of development and implementation of different programmes and policies. Administrators often blame too high level of expectations or too high inertness of society and economics as the reasons for failures.

Meanwhile in our field researches (conducted by the Department of Local Administration) we've seen the problems of collecting relevant information, caused not only by the official statistics system shortcomings, but by the lack of mechanisms of ranking the threats and needs at the state and local power at all.

From our point of view, failures in public administration are often caused by the unreliable information and poor tools of its analysis, which, in fact, can compensate each other.

I'm not a computer student, but it is supposed that in this particular case interdisciplinary approach is the most appropriate.

In order to verify this hypothesis and develop effective mechanism of collecting and analyzing data the following research would be carry out.

2 Overview

- What tools (formal and informal) do local authorities use in order to fill the gaps in official statistics' data?
- Do the mechanisms of collecting information correlate with the essence of actions or its type? (Could some interconnections like "statistics are used for long-turn programmes, data from the mass media – for day-to-day management" be revealed or not?)
- What modern tools of collecting and analyzing information offered by IT market could be more reasonable and effective for satisfying local self-government informational demand?
- How could relevance and validity of information currently used by municipal authorities be evaluated? What are the reasons for misrepresentation?
- How could quantitative and qualitative data collected through statistics and conceptual analysis of the mass media be combined within the bound of one model (system)?

3 Planned Methodology

The object of the investigation is the departments of municipal governments and specialized agencies responsible for the analysis of information space and the collection of socio-economic situation indicators (providing the basis for administrative decisions).

The subject of the research is governance relations arising in the process of information-analytical support of the local authorities.

The structural functional analysis (approach) is used in order to achieve objectives.

In accordance with the objectives the following research methods will be used:

Theoretical: forming hypotheses about the quality and effectiveness of the tools currently used by local authorities to collect and analyze information and the ways of its improving; scientific modeling of information flows in the separate municipalities and the usage of their content as a foundation for developing programmes, policies, plans, projects and so on.

Empirical: keeping under observation (descriptive research method) internal organization of the information-analytical departments; an experiment in particular municipality to access the effectiveness of proposed model of integrated socio-economic indicators and the needs of social community collection and analysis; scientific research (search, application and development), aimed at measuring the topicality of investigation for municipal authorities.

The information collected through the field research would be also analyzed with statistical program package (Stata) in order to reveal interrelations. The conceptual analysis would be carried out through the system Gitika (partly, through the website http://www.gitika.ru/, restricted access).

4 Progress made so far

To date the following steps are fulfilled:

- developing the methodology of field research (choosing of municipal units, way of collecting data, obtaining funds for survey);
- reviewing the literature according to the topic (ongoing);
- developing the system of evaluation urgent needs of citizens based on the analysis of the mass-media (ongoing).

5 Future plan

It is set up in compliance with the objectives. This year the field research would be fulfilled, after which, possibly, the methodology will require some correction. Anyway, developing plan for creating the model would be substantiate only after careful investigation of current systems of collecting and analyzing data.

References

1. Kordonskii S.G. Threat as an institution of resource management (in press).
2. Kordonskii S.G., Bardin V.V. On The Search For Information In The Totality of Texts That Represents The Picture Of The World. - Washington, RusGenProekt, 2010. - 62.
3. Pacesila M., Profiroiu A. Recent Evolutions Concerning the study of Public Policy. / / Administratie Si Management Public. 7/2006. pp. 149 - 156
4. Content Analysis: A Methodology for Structuring and Analyzing Written Material. United States General Accounting Office. Transfer Paper 10.1.3 March 1989
5. Vlieger E., Leydesdorff L. Content Analysis and the Measurement of Meaning: The Visualization of Frames in Collections of Messages. / / The Public Journal of Semiotics III (I), June 2011. pp. 28 - 50
6. Bessonov. V.A. On the problems of development of the Russian statistics / / IVF. 2012. Number 3. pp. 35 - 49
7. Baranov E.F. Russian statistics: achievements and challenges / / ECO. 2012. Number 3. pp. 23 - 34
8. Franzosi R. Content Analysis: Objective, Systematic, and Quantitative Description of Content. / / Content Analysis. Benchmarks in Social Research Methods series (Quantitative Applications in the Social Sciences). 4 vols. Thousand Oaks, CA: Sage. 2007. pp. xxi - 1
9. Media research: methodology, approaches, methods: a teaching aid. - Moscow: Mosk. University, Faculty of Journalism of Moscow State University, 2011. - 236 p.
10. Language media as an object of interdisciplinary research. Under. Ed. Volodin M.N. - Moscow: Moscow State University Press, 2003. - 320.
11. Racer S.A. Fundamentals of textual study. 2nd Ed. - L: "Education", 1978. - 176.
12. Afanasyev D.G. Conceptualization and measurement of threats and risks in the development of the region's anti-crisis program
13. Report, "The organization of strategic monitoring and auditing in the public programs and projects" (made in accordance with the plan of the expert and analytical work of the federal government agency "Analytical Centre of the Government of the Russian Federation" in

2011) / / Federal Government Agency "RESEARCH CENTRE GOVERNMENT OF THE RUSSIAN FEDERATION " http://ac.gov.ru/files/audit-doklad.pdf

14. Ex-ante Evaluation. A Practical Guide for Preparing Proposals for Expenditure Programmes. European Commission, 10 December 2001. http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/ex_ante_guide_2001_en.pdf
15. Information-analytical support of legislative activity: Issues and Experiences / / Council of the Federal Assembly of the Russian Federation. Analytical Bulletin number 2 (158). - M., 2002. - 114 p.
16. The integrated system of information and analytical support of the executive authorities of St. Petersburg. http://kis.gov.spb.ru/projects/project-sistema-nif-vzaimodeistia/
17. Analysis and evaluation of government programs and industrial policies: reader to learn. discipline / Tsygankov D.B.; Smirnov M.V.; Settles A. - Moscow: Higher School of Economics, 2006. - 252 p.
18. Yadov V.A. The strategy of the survey. Description, explanation, understanding of social reality. - M.: "Dobrosvet", 1998.
19. Pashintseva N.I. Modern problems of statistics in regions and municipalities / / Problems of Statistics, 2006. Number 12. - With. 5 - 11
20. Gurinovich A.G. Guidelines of information and analytical support of local government
21. Professiograme of specialists working in information and analytical service authority. The expert study. / / Research Center "Analytic", Ekaterinburg. 2011. http://www.rc-analitik.ru/file/% 7B2665990d-6d38-4b08-bf92-d7c33d5b5490% 7D
22. Semenov A., Korsun M. Content analysis of the media: issues and experience/ Under. Ed. Mansurov V.A. - Moscow: Institute of Sociology, 2010. - 324 p.
23. Webster D. Unemployment: How Official Statistics Distort Analysis and Policy, and Why. http://www.radstats.org.uk/no079/webster.htm
24. Shadow Government Statistics. Analysis Behind and Beyond Government Economic Reporting. http://www.shadowstats.com/
25. Proposed Standards and Guidelines for Statistical Surveys. White House. http://www.whitehouse.gov/sites/default/files/omb/inforeg/proposed_standards_for_statistical_surveys.pdf
26. Guidance on the development and implementation of state programs in the Russian Federation. Approved by order of the Russian Ministry of Economic Development 22/12/10. № 670
27. Methods of foundation of the rural areas sustainable development: monograph. / Under. Ed. Frolov V. - St.: St. Petersburg. State. archit.-engin. un-ty, 2011. - 464.
28. Fetisov G.G., Oreshin V.P. Regional Economics and Management: A Textbook. - Moscow: INFRA-M, 2006. - 416.

Development of Methods for a Problem in Multicriteria Stratification

Mikhail Orlov

1st year PhD student,
Higher School of Economics, Russian Federation

Supervisor Prof. Boris Mirkin

Abstract. In decision making and data analysis it may be useful to order items or set them into homogeneous groups over multiple criteria. Supervised classification of items into several ordered classes is often called multicriteria sorting. There is no common terminology in the case of unsupervised ordered classification. We refer to this problem as multicriteria stratification. This problem is closely related to multicriteria ranking problem as ordered items can be easily broken into ordered groups of items. But there is no certain rule how to get this partition and usually it is left to the decision maker. There are also some stratification methods employing information about preferences of the decision maker. In fact, there is no entirely automatic stratification method. The focus of this research is to fill in the niche by developing stratification methods that would be able to rank items, find optimal strata and criteria weights by using only multicriteria data provided. This can be achieved by using an optimization criterion based on the “geometrical” meaning of goodness of stratification. We propose and explore optimization procedures for the criterion. This can turn stratification into a useful and efficient tool for the analysis of multicriteria data in such fields as information retrieval, recommender systems and risk management.

1 Motivation

In the literature lack of attention is paid to the multi criteria stratification or ordered clustering problem. It is often assumed that stratification problem can be solved by means of the methods from the related fields such as multicriteria ranking or decision theory. One of the well-studied approaches for multicriteria ranking is weighed sum of criteria. For the purpose of rank aggregation multiple criteria values are summed up into a single one. Weights can be received from experts or calculated in some other way. In the paper [Ng 2007; Ramanathan 2006] criteria weighted sum is used for so called ABC–classification task for the case of multiple criteria. In this case weights are found by formulating and solving linear optimization task. In [Sun at al 2009] authors developed a method for simultaneous evaluation both ranks and weights for the task of conferences and their participants ranking.

There exists a number of rank aggregation methods from the public choice theory [Aizerman, Aleskerov 1995; Mirkin 1979]. These methods allow combining several rankings by means of some aggregation rules.

Various multicriteria stratification methods have been proposed in the decision theory. These methods exploit decision maker's (DM) preferences. In [DeSmet, Montano, Guzman 2004] extended version of classical k-means algorithm is developed. The proposed k-means modification is based on a new kind of distance between items considering structure of DM's preferences. An approach based on pair-wise comparison of items is studied in [Nemery, DeSmet 2005].

Aforementioned methods require some information from DM. In multicriteria ranking items have to be manually appointed on corresponding stratum. In decision theory one has to reveal decision maker's preferences. Therefore, both stratification approaches are not entirely automatic. This make difficult to use them for large sets of items evaluated on a large number of criteria.

The main motivation of the research is developing of entirely automatic methods for simultaneous criteria weighting, ranking and stratification. This can make stratification a convenient and efficient tool for data analysis and decision making including the situations where processing of large amount of data is needed. Examples of applications may be found in various fields such as:

- Information retrieval (web services ranking and clustering) [Skoutas et. al. 2010];
- Risk management (country risk problem) [De Smet, Gilbart 2001];
- Marketing (multicriteria ABC-classification) [Ng 2007; Ramanathan 2006].

An interesting "geometrical" criterion for automatic determination of both weights of criteria and strata has been proposed by B. Mirkin (2011); in some experimental evaluations it showed promising results (Orlov 2012).

2 Research questions

Here is a list of some research issues we are going to address:

- How the geometrical criterion is related to other data summarization criteria?
- How synthetic stratified data can be generated to correspond to real world multicriteria decision making situations?
- Is there any difference between the results of newly proposed stratification methods and those described in the literature methods at different "shapes" of strata?

3 (Planned or ongoing) Methodology

The planned methodology of the research is related to methods and models of multicriteria ranking, decision theory, optimization (evolutionary, linear and quadratic programming), data clustering, artificial neural networks (auto-encoders) and design of experiment.

4 Progress made so far

By this moment, I have developed the following:

1) An evolutionary minimization procedure for the geometrical stratification criterion is developed and implemented;

2) A quadratic programming based algorithm for the geometrical stratification criterion is proposed and implemented;

3) A neural network representation of stratification criterion is developed. An algorithm for learning the network parameters is implemented;

4) Generative stratified data models for different types of stratified data are developed and implemented;

5) Most popular known multicriteria ranking and ordered clustering methods are implemented;

6) Preliminary experimental study and comparison of the proposed and existed methods on synthetic datasets is performed.

5 Future plan

Planned activities towards the further research are:

- Investigation of properties of the geometrical criterion of stratification and its comparison to other data summarization criteria;
- Further development of generative models of synthetic “stratified” data and experimental study of the existed and proposed methods on the generated data;
- Extensive experimental study of the developed and existing stratification methods for the real world applications.

References

1. Aizerman M., Aleskerov F. Theory of Choice, Elsevier, North-Holland, 1995, 314 pp.
2. Aleskerov F., Ersel H., Yolalan R. (2004) Multicriteria ranking approach for evaluating bank branch performance // International Journal of Information Technology & Decision Making. Vol. 3, No. 2, 321-335
3. De Smet Y., Montano Guzman L. (2004) Towards multicriteria clustering: an extension of the k-means algorithm // European Journal of Operational Research. 158(2). pp. 390-398
4. DeSmet Y., Gilbart F. (2001) A class definition method for country risk problems.Technical report IS-MG 2001/13. 2001.
5. Mirkin B. (1979) Group Choice.Translated by Yelena Oliker. Edited and introduced by. Peter C. Fishburn. Washington, D. C.: V. H. Winston, pp. xxx + 252
6. Mirkin B. (2011) A data approximation criterion for multicriteria stratification (Personal communication).
7. Nemery Ph., De Smet Y. (2005) Multicriteria ordered clustering. Technical Report TR/SMG/2005-003. Universite Libre de Bruxelles. 2005.
8. Ng W.L. (2007) A simple classifier for multiple criteria ABC analysis // European Journal of Operational Research 177. pp. 344–353.

9. Orlov M. (2012) Multi-criteria stratification methods and their experimental comparison, A MSc diploma project, Division of Applied Mathematics and Informatics, NRU Higher School of Economics,
10. Ramanathan R. (2006) Inventory classification with multiple criteria using weighted linear optimization // Computers and Operations Research. 33. pp. 695-700.
11. Skoutas D., Sacharidis D., Simitsis A., Sellis T. (2010) Ranking and Clustering Web Services Using Multicriteria Dominance Relationships. IEEE T. Services Computing 3(3): 163-177
12. Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T. (2009) RankClus: integrating clustering with ranking for heterogeneous information network analysis // Proc. EDBT. 2009. pp. 565-576.
13. Zopounidis C., Doumpos M. (2002) Multicriteria classification and sorting methods: A literature review // European Journal of Operational Research. 138. 229–246

Applications of Textual Entailment

Partha Pakray*+, Sivaji Bandyopadhyay+, Alexander Gelbukh#

*4th Year PhD Student
+ Computer Science and Engineering Department,
Jadavpur University, Kolkata, India
# Center for Computing Research, National Polytechnic Institute,
Mexico City, Mexico
parthapakray@gmail.com,
sbandyopadhyay@cse.jdvu.ac.in, gelbukh@gelbukh.com,

Abstract. Given two texts called T (Text) and H (Hypothesis), Textual Entailment task consists in deciding whether or not the meaning of H can be logically inferred from that of T. We have development text entailment techniques. Here, we present the Answer Validation (AV) System and Question Answering (QA) System with the help of Textual Entailment (TE) techniques. Our textual entailment technique applied into AV system and QA system. The TE engine is based on Named Entity Recognition (NER), Question- Answer Type Analysis, Chunk Boundary, Lexical Similarity, Syntactic Similarity and Semantic similarity that are integrated using a voting technique. For AV system we have used Answer Validation Exercise[1] datasets and QA system we have used QA4MRE@CLEF 2012[2] datasets. For AV system, we have combined the question and the answer into the H, and consider the Supporting Text as T to identify the entailment relation as either “VALIDATED” or “REJECTED”. For QA system, we first combine the question and each answer option to form the Hypothesis (H). Stop words are removed from each H and query words are identified to retrieve the most relevant sentences from the associated document using Lucene. Relevant sentences are retrieved from the associated document based on the TF-IDF of the matching query words along with n-gram overlap of the sentence with the H. Each retrieved sentence defines the Text T. Each T-H pair is assigned a ranking score that works on textual entailment principle. For the AV system (on the AVE 2008 English test set), we obtained precision of 70% and F-score of 68%, and for the QA system score c@1 is 0.65.

Keywords: Textual Entailment (TE), Answer Validation Exercise (AVE), Question Answering (QA).

1 Research Motivation

The recognition of textual entailment is one of the recent challenges and most demanding of the Natural Language Processing (NLP) domain. Recognition of textual

[1] http://nlp.uned.es/clef-qa/ave/
[2] http://celct.fbk.eu/ResPubliQA/index.php

entailment is such a field of Natural Language Processing (NLP) where all other specialized branches are emerged. Due to this aspect system developed for entailment must consider the other research study of NLP. Entailment can come out from its own field of study and can be linked with other tasks of NLP such as: Summarization (SUM): a summary should be entailed by the text; Paraphrases (PP) can be seen as mutual entailment between a text T and a hypothesis H; Information Extraction (IE): the extracted information should also be entailed by the text; Question Answering (QA) the answer obtained for one question after the Information Retrieval (IR) process must be entailed by the supporting snippet of text; Machine Translation (MT) translation should be semantically equivalent to the gold standard translation, i.e., must entail each other. In NLP, same meaning can be expressed by, or inferred from, different texts. All the above mentioned tasks require a model that recognizes a particular target meaning that can be inferred from different text variants. Entailment can be defined as a relation that holds between two language expressions (i.e. a text T and a hypothesis H) if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T. The relation is directional as the meaning of one expression can entail the meaning of the other, but not true for the vice versa. The Recognizing Textual Entailment (RTE) Challenge introduces a generic task that combines the semantic inferences required across NLP applications. Evolution of this task has been done through different RTE Challenges. Every challenge adds some new flavour to the task compared to its predecessors.

2 Methodology

We have developed textual entailment system in mono-lingual and cross-lingual scenario. We are applying our textual entailment technique in Answer Validation and Question Answering system. For textual entailment we have developed both rule based and machine learning based systems that consider the lexical, syntactic, semantic features. We mainly target to establish the entailment relation based on lexical, syntactic and semantic similarity measures. We have measured Named Entity (NE), Parts-of-Speech (POS), WordNet Similarity, Chunk Similarity, N-Gram Similarity, Lexical Distance Similarity, Syntactic Similarity, Semantic Similarity and other to decide the entailment relation through rule based and machine learning based approaches. We have focused the entailment problem as a classification problem. We have experienced to participate in different evaluation tacks that mainly concentrated on issues of

i) Two-way Entailment
ii) Multi-way Entailment
iii) Cross Lingual Entailment Scenario (e.g. Two-way, Multi-way)

In Recognizing Textual Entailment (RTE)[3] of Text Analysis Conference (TAC) at its 5th edition (2009) Main Task we proposed a lexical based system [1]. In RTE of TAC

[3] http://www.nist.gov/tac/

at its 6th edition (2010) Main Task we proposed a lexical based and syntactic similarity system [2]. In RTE of TAC at its 7th edition (2011) we proposed a lexical and syntactic based system [3] with anaphora resolution as a preprocessing task. Recognizing Inference in Text (RITE)[4] task in NTCIR-9 (2011) has proposed new direction of multiclass entailment in mono-lingual (Asian Languages) Scenario. We have developed a system named "*A Textual Entailment System using Web based Machine Translation System*" [4] to address this problem. It is a rule based system that measures N-Gram and text similarity property of the entailment pairs. Later we have developed a machine learning based system that has experimented over the RITE dataset. Semantic Evaluation Exercises (SemEval-2012)[5] has proposed multiclass textual entailment in a cross-lingual scenario. Machine translation is an essential integrated framework to address the problem of cross-linguality. We have developed a system named "*Language Independent Cross-lingual Textual Entailment System*" [5] that also use machine translation to address the cross lingual problem. We have developed another system for Semantic Evaluation Exercises (SemEval-2012) named "*Multi-grade Classification of Semantic Similarity between Text Pair*" [6] that is based on similarity score of entailment pairs instead of entailment labels. We have used the Universal Networking Language (UNL)[6] to identify the semantic features. The development of a UNL based textual entailment system [7] that compares the UNL relations in both the text and the hypothesis. So, we first develop the different textual entailment system and participated different evaluation track. Now we applied that entailment system to Answer Validation system and Question Answering system.

We present an Answer Validation System (AV) [8] based on Textual Entailment. We first combine the question and the answer into Hypothesis (H) and the Supporting Text as Text (T) to check the entailment relation as either "VALIDATED" or "REJECTED". We have applied the Textual Entailment technique to AV system to detect the relation of "VALIDATED" or "REJECTED". The architecture of the proposed Answer Validation (AV) system is described in Figure 1.

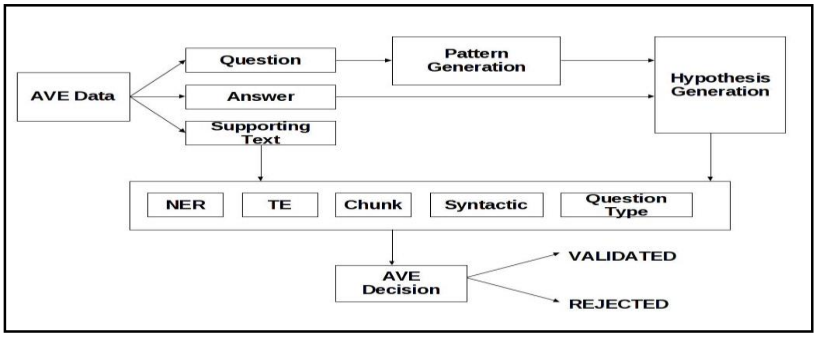

Figure 1 Answer Validation System

[4] http://artigas.lti.cs.cmu.edu/rite/Main_Page
[5] http://www.cs.york.ac.uk/semeval-2012/
[6] www.undl.org/unlsys/unl/UNLSpecs33.pdf

A number of answer validation modules has been developed based on Textual Entailment, Named Entity Recognition, Question-Answer type analysis, chunk boundary module and syntactic similarity module. These answer validation modules have been integrated using voting technique. We combine the question and the answer into Hypothesis (H) and the Supporting Text as Text (T) to check the entailment relation as either "VALIDATED" or "REJECTED". The details of the system are presented in [8]. For AV system we have used AVE @CLEF 2008 datasets. We have trained our system by AVE development dataset and tested on AVE 2008 test data. The recall, precision and f-measure values on the test data obtained over correct answers are shown in Table 1.

Table 1. Experiment Result for AV

	AVE Development Set	**AVE Test Set**
"VALIDATED" in Data Set	21	79
"VALIDATED" in the proposed AV system	32	76
"VALIDATED" match	18	54
Precision	0.56	0.70
Recall	0.85	0.68
F-score	0.68	0.68

We have also applied the textual entailment technique in Question Answering (QA) Track (e.g. QA4MRE) in CLEF 2011 [9] and CLEF 2012 [10]. In QA4MRE@CLEF2011[7] our system showed the first place best result out of 10 participating systems for the task with 0.57 and in QA4MRE@CLEF2012[8] the highest score was obtained by our team with 0.65. The architecture of QA system is described in Figure 2. Proposed architecture is made up of four main modules along with knowledgebase i.e Document Processing, Validate Factor Generator, Inference Score Module, Answer Option Selection Module. We have applied our TE technique in Validate Factor Generator Module and for Answer selection ranking. The details of the system are presented in [10].

[7] http://clef2011.org/resources/proceedings/Overview_QA4MRE_Clef2011.pdf

[8] http://www.clef-initiative.eu/documents/71612/c076dd78-e36b-40d9-a6c8-fed4b7aa0b3d

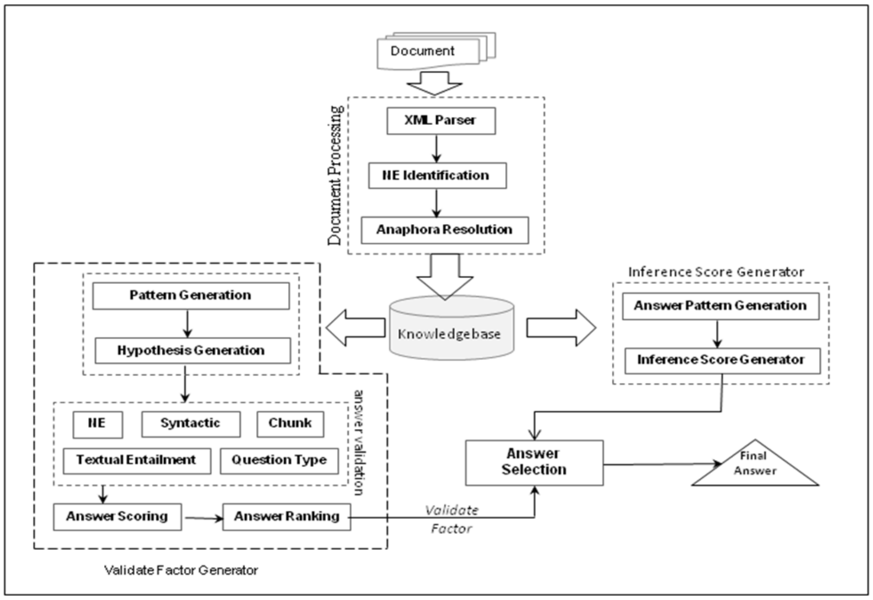

Figure 2 QA System

We have used QA4MRE@CLEF 2012 datasets for our experiment. Overall *accuracy* of that system is 0.53 and Overall *c@1 measure* is 0.65.

3 Future Plan

The study we have done so far on textual entailment mainly focused on lexical, syntactic and semantic approaches. We want to include more semantic approaches into our study in future. Conceptual Graph representation of sentences can be useful to gather the semantic role knowledge. Graph matching will be supportive to decide the entailment class. In case of a text hypothesis pair we represent them through the conceptual graph to get an idea of semantic knowledge. Then we simply match the nodes of the graphs to finally decide the entailment. We are trying to apply our entailment system to

i. Summarization evaluation
ii. Machine Translation evaluation

Acknowledgements. The first author acknowledges his Doctoral supervisors Professor Sivaji Bandyopadhyay, India and Prof. Alexander Gelbukh, Mexico and the support of the DST India-CONACYT Mexico project "*Answer Validation through Textual Entailment*".

References

1. Partha Pakray, Sivaji Bandyopadhyay, Alexander Gelbukh, "Lexical based two-way RTE System at RTE-5", Text Analysis Conference Recognizing Textual Entailment Track Notebook, 2009.
2. Partha Pakray, Santanu Pal, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh "JU_CSE_TAC: Textual Entailment Recognition System at TAC RTE-6", Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook, 2010. [2010]
3. Partha Pakray, Snehasis Neogi, Pinaki Bhaskar, Soujanya Poria, Sivaji Bandyopadhyay, Alexander Gelbukh, "A Textual Entailment System using Anaphora Resolution", Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook, November 14-15, 2011.
4. Partha Pakray, Snehasis Neogi, Sivaji Bandyopadhyay, Alexander Gelbukh, "A Textual Entailment System using Web based Machine Translation System.", RITE competition: Recognizing Inference in TExt@NTCIR9. December 6-9, 2011 [2011]
5. Snehasis Neogi, Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh: "JU_CSE_NLP: Language Independent Cross-lingual Textual Entailment System", (*SEM) First Joint Conference on Lexical and Computational Semantics, Collocated with NAACL-HLT 2012, June 7-8, Montreal, Canada. [2012]
6. Snehasis Neogi, Partha Pakray, Sivaji Bandyopadhyay, Alexander Gelbukh: "JU_CSE_NLP: Multi-grade Classification of Semantic Similarity between Text Pair", (*SEM) First Joint Conference on Lexical and Computational Semantics, Collocated with NAACL-HLT 2012, June 7-8, Montreal, Canada. [2012]
7. Partha Pakray, Utsab Barman, Sivaji Bandyopadhyay and Alexander Gelbukh, "Semantic Answer Validation using Universal Networking Language", In International Journal of Computer Science and Information Technologies (IJCSIT), ISSN 0975 - 9646,PP. : 4927 - 4932, VOLUME 3 ISSUE 4 July- August 2012.
8. Partha Pakray, Alexander Gelbukh and Sivaji Bandyopadhyay, "Answer Validation using Textual Entailment", 12th International Conference on Intelligent Text Processing and Computational Linguistics, 2011, pp. 359-364 [2011]
9. Partha Pakray, Pinaki Bhaskar, Somnath Banerjee, Bidhan Chandra Pal, Sivaji Bandyopadhyay and Alexander Gelbukh, "A Hybrid Question Answering System based on Information Retrieval and Answer Validation", CLEF 2011 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE). [2011]
10. Pinaki Bhaskar, Partha Pakray, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay and Alexander Gelbukh, "Question Answering System for QA4MRE@CLEF 2012", CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE). [2012]

A Study to Develop Appropriate Tools and Methodologies for Searching and Vetting of Originality in New Copyrights, Patents

Ranjeet Kumar (RS-98)

2nd year PhD student

Indian Institute of Information Technology Allahabad, India

Supervisor: Prof. R.C.Tripathi

Abstract. As the world is growing economically, the Intellectual Property profile of any country/corporate is likely to grow progressively. IPRs are the bottom rocks for global operations by multinational companies/corporates. Almost in all the corporates, Academic world and Research Labs, new knowledge is being advanced by researches which are getting added attention. New Copyrights and Patents profile of all above domains is growing globally somewhat by 10% to 15% annually. Copyrights issues are getting more focused now days in Indian Academic R&D labs and individuals - circles and the copyright violations like plagiary in research papers and thesis etc. is now being viewed as very serious cognizable offense. The valuation of research and the continued pursuit for seeking the best research output of an Academic institution or any Global research lab demands its patents and Copyrights to play a very critical and important role in the present competitive world. So only above two IPR's have been chosen for present Ph.D. work looking into their prime importance to academic institutes and industry of the ICT Sector. The Main objective of the present work is finally to provide the relevant search report based on the inputs like text, audio, images, video, symbols, diagrams, logo etc. The use of Information Retrieval (IR) techniques for retrieving the data from the internet as well as from the local databases/repositories will be focused for the searching and matching to arrive at the originality of the given input. The latest techniques are targeted to be used to fetch and provide efficient and most relevant data in above pursuit.

1 Motivation

Towards vetting of copyrights of a research paper, it is observed that, reference citation is the major issue for the researches and the citation of the research work related to the target topic must be relevant and somehow related. In this regard, the relevancy check of the references cited in the research paper is essential. In the case of copyrights issues in the textual content of the research papers and issues of new ideas of the patents, the exact retrieval of the relevant material is an arduous task. The information in the form of the textual as well as other contents and its retrieval will be the main area of the present research topic. The research in this regard may use vividly many information retrieval techniques as well as graphics and image processing. The exact textual content retrieval and matching the same with the query document is the main motivation of the topic.

2 Research questions

- To study and develop the methodologies for the retrieval of the relevant data from the internet as well as local databases for the given text given is proposed to be investigated. This is at the crux of the issues of the copyright like references cited in the research papers, free websites used for the content and in-house matter Para phrased. The efficient techniques would be used for the data retrieval problem and will be tested on various aspects to verify the results to enhance the performance of the search techniques to be developed.
- To develop and enhance the query and semantic search results in the existing database search results for the in-house plagiarism detection tool will also be focused. The technique used in the existing software will be enhanced in terms of performance and the accuracy of the results.
- To create a database for 7-8 US Patent Classes using the US Patent website. To create a user option to choose a patent search on the basis of Keywords, US Class or assignee. Apply the technique of information retrieval for the same and give the results. The searching of the related documents in the database is aimed to be semantic to list the similar found patents with the details.

3 Methodology (Planned or ongoing)

Copyrights are one of the most important IPR’s for the Academia and the researchers community. In the academic research, the publications of the Books, Dissertations, Research Papers, Thesis, Softwares and other scientific and technical articles are prone to the Plagiarism issue. It is very critical challenge for the researcher community to prevent these cases in the publications and other copyright matter in the academia. There are different ways to protect the data from plagiarism issues online like “Digital Rights Management (DRM)” or in the hard copy of the book and unpublished data.

In the proposed methodology of the copyright issues in the academia, plagiarism research is also going on by some agencies. So many methodologies and techniques have been proposed for the textual plagiarism as evident from the internet literature. In the proposed process for the same, in figure 1 shows the future plan work flow of the proposed methodology. It consists of three different modules.

A) Search on the internet for the relevant data on the basis of query data,
B) Search the references cited in the research papers, dissertation or thesis and then check the relevancy of the content as well as references,
C) Search in the last 4 years of data in the same field and then combine them all for the final output of the data on the basis of query data.

Patents prior art search methodology for a patent consists of different aspects and modules for the relevant search such as full body patent search, semantic search, followed for continuous and content search to achieve the higher relevancy of the input documents. Use will be made of the US ICT patent classes for the related search for the query data. After search, each of the retrieved data will be interviewed by user with a set of pre-designed queries and on some casual information searched on those queries during the search session. The main results and their implications for developing future patent retrieval systems will be generated after all processes feedback is obtained from the users.

4 Progress made so far

In the proposed methodology, the work has been divided in the several parts and the work flow accordingly is being maintained. In the starting period of the research, the prior art or literature survey has been completed regarding the problem formulations and the exact position of the research in this regards. A Database has been created with 2,000 of the research papers of different fields the local repository with separate folders for different domain research respectively has been created. A Local repository of the 2, 000 US Patents has been created and maintained for the testing purposes of the developed methodology for the similar patent document retrieval.

5. Future plan

In the proposed process, the references will be examined for the relevancy of the content which is used in the research papers and in others. References should be in the relevant order not only for save of the number or count. The matter of the relevancy of the written content will only be referenced and no other irrelevant referencing should be there. In the third part of the search, the database created within the 4 years of data of research papers, dissertations and thesis of the relevant fields will be focused. These set of searches will be performed and then on the global internet search will be performed and then finally all the results will be combined generate the final output for the user query data.

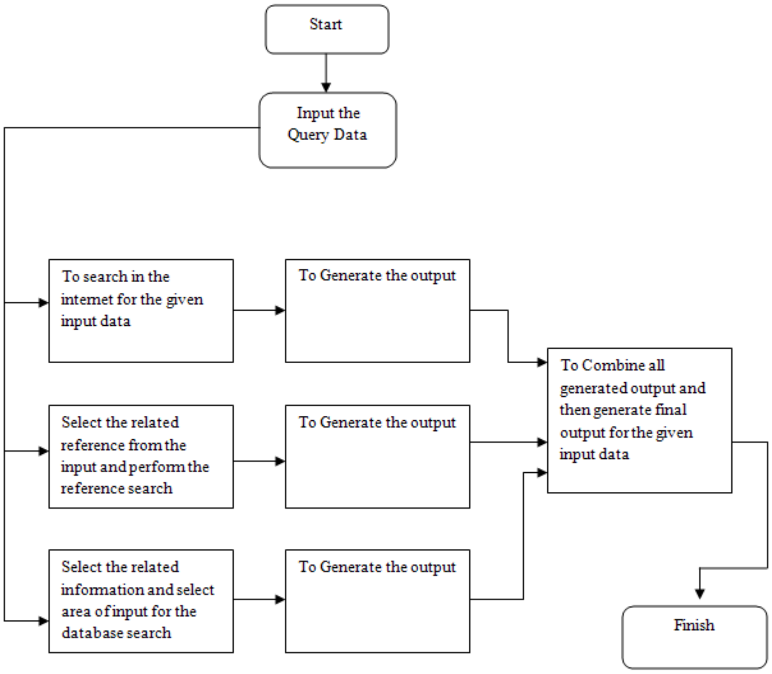
Start
Input the Query Data
To search in the internet for the given input data
To Generate the output
Select the related reference from the input and perform the reference search
To Generate the output
Select the related information and select area of input for the database search
To Generate the output
To Combine all generated output and then generate final output for the given input data
Finish

Figure 1. The proposed work flow diagram for the copyright search process.

Figure 1 shows the proposed techniques and methodologies to be developed for the different level of the modules in the system and finally generate the efficient and relevant output which will cover all the possible ways to protect the copyright issue in the particular case.

Patents:

Patent prior art search for the relevancy of the new textual expression or new innovation is the most challenging issue in the field of Intellectual Property Rights (IPR’s). The researches have been tried for many techniques of the retrieval of the most relevant text for copyrights and the patents for the searching of the patentability or novelty search of the new innovations. Some of them are online query based search whereas some of others are based on databases. Different parameters have been defined for the patent search because of the complex format of the patents. Well known Patent Searches are outlined as below-

i) Abstract Based Search: The patent search on the given abstract is, used as a query to search and generate by the users the patents having almost same abstract. But the patent abstract has very brief information and is written in the manner which uses new and twisted languages. Abstract has no details of the actual processes of methodologies of the innovations, so in the prior art search based on abstract, the results may not be satisfactory in nature.

ii) Keyword Based Search: The keyword search is the process in which the query generated by the users enables to search the relevant data. In the keyword search there is possibility to search large scale of data on the different sections of the patent like abstract, detailed descriptions, and in brief summary of the innovation.

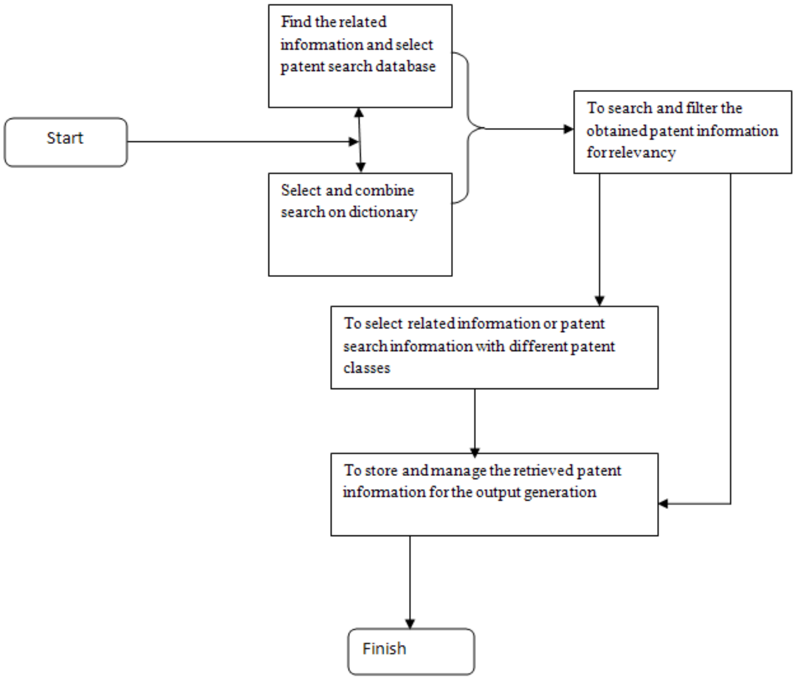
Find the related information and select patent search database
Start
Select and combine search on dictionary
To search and filter the obtained patent information for relevancy
To select related information or patent search information with different patent classes
To store and manage the retrieved patent information for the output generation
Finish

Figure 2. Proposed Patent Prior Art Search Work flow diagram

References

[1] H. Zhang and T. W. S. Chow, “A coarse-to-fine framework to efficiently thwart plagiarism,” Pattern Recog., vol. 44, pp. 471–487, 2011.

[2] M. Roig, *Avoiding Plagiarism, Self-Plagiarism, and Other Questionable Writing Practices: A Guide to Ethical Writing*.NewYork: St. Johns Univ. Press, 2006.

[3] J. Bloch, “Academic writing and plagiarism: A linguistic analysis,” *English for Specific Purposes*, vol. 28, pp. 282–285, 2009.

[4] I. Anderson, “Avoiding plagiarism in academic writing,” *Nurs. Standard*, vol. 23, no. 18, pp. 35–37, 2009.

[5] K. R. Rao, “Plagiarism, a scourge,” *Current Sci.*, vol. 94, pp. 581–586, 2008.

[6] Delvin, M. Plagiarism detection software: how effective is it? Assessing Learning in Australian Universities, 2002. Available at: http://www.cshe.unimelb.edu.au/ assessinglearning/docs/PlagSoftware.pdf

[7] iParadigms. Plagiarism Prevention: Stop plagiarism now… with Turnitin®. Product datasheet, 2006. Available online at: http://turnitin.com/static/pdf/datasheet_plagiarism.pdf

[8] iParadigms, LLC. Turnitin. Plagiarism prevention engine. Available online at: http://www.turnitin.com

[9] Lancaster T., F. Culwin. A review of electronic services for plagiarism detection in student submissions. Paper presented at 8th Annual Conference on the Teaching of Computing, Edinburgh, 2000. Available at: http://www.ics.heacademy.ac.uk/events/ presentations/317_Culwin.pdf

[10] Lancaster T., F. Culwin. A visual argument for plagiarism detection using word pairs. Paper presented at Plagiarism: Prevention, Practice and Policy Conference 2004.

[11] Lancaster, T., F. Culwin. Classifications of Plagiarism Detection Engines. ITALICS Vol. 4 (2), 2005.

[12] Maurer, H., F. Kappe, B. Zaka. Plagiarism – A Survey. Journal of Universal Computer Sciences, vol. 12, no. 8, pp. 1050 – 1084, 2006.

[13] Neill, C.J., G. Shanmuganthan. A Web – enabled plagiarism detection tool. IT Professional, vol. 6, issue 5, pp. 19 – 23, 2004.

[14] The University of Sydney Teaching and Learning Committee. Plagiarism detection software report. Draft One, 2003.

[15] Nanba, H. 2007. Query Expansion using an Automatically Constructed Thesaurus. Proceedings of the 6th NTCIR Workshop, pp.414-419.

[16] Larkey, L.S. (1999). A patent search and classification system. Proceedings of the fourth ACM conference on Digital libraries. pp. 179 – 187

[17] Inoue, N., Matsumoto, K., Hoashi, K., Hashimoto, K. (2000). Patent Retrieval System Using Document Filtering Techniques. Proceedings of the Workshop on Patent Retrieval, ACM SIGIR.

[18] Lim, S.-S., Jung, S.-W., Kwon H.-C. (2004). Improving patent retrieval system using ontology. Industrial Electronics Society, 2004. IECON 2004. 30th Annual Conference of IEEE.

[19] Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., Oshio, T. (2005). Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing (TALIP), Volume 4 Issue 2.

[20] Anthony Trippe and Ian Ruthven, Evaluating Real Patent Retrieval Effectiveness, Current Challenges in Patent Information Retrieval, The Information Retrieval Series, 2011, Volume 29, part 2, 125-143, DOI 10.1007/978-3-642-19231-9_6

[21] John I.Tait and Barou Diallo, Future Patent Search, Current Challenges in Patent Information Retrieval, The Information Retrieval Series, 2011, Volume 29, part 5, 389-407, DOI 10.1007/978-3-642-19231-9_20

[22] Ferdinand De Laet "Patentless" center for Patent Information retrieval at ARIPO, World Patent Information, Science Direct, Volume 28, Issue 3, September 2006, Pages 212-214

[23] [26] Sergey Butakov, and Vladislav Scherbinin, The toolbox for local and global plagiarism detection, Computers & Education, Elsevier Publications, Volume 52, Issue 4, May 2009, Pages 781-788.

[24]R ebecca Moore Howard, Understanding " Internet Plagiarism", Computers & Composition, Elsevier Publications, Volume 24, Issue 1, 2007, Pages 3-15

[25] Rudi bekkers, Rene Bongard, and Alessandro Nuvolari, An empirical study on the determinants of essential patent claims in compatibility standards, Research Policy, Elsevier Publications, Article in Press, 2011.

[26] Campbell, R. S. Patent trends as a technological forecasting tool. World Patent Information, 5(3), 137–143 (1983).