PROPOSAL

of the Diploma Project entitled (working title(s)):

A Unified Perspective on Context-Sensitive Content

**Unified Context-Behavioral Ad Recommendation System**

ConSys – Ad Recommendation System

|  |  |
| --- | --- |
| Author: | **Florin Cristian MACICASAN** |
| Advisor: | **Prof. Dr. Eng. Rodica POTOLEA**  **Conf. Dr. Eng. Mihaela DINSOREANU** |

# Project Purpose and Objectives

The purpose of our project is to define and construct a system capable of providing a context-sensitive recommendation to a user based on the triple similarity relation between [the currently active context](#currentactive), an advertisement and the user’s historical information.

Our main objectives are:

* Extracting the part of the context with the [highest descriptive value](#descriptivevalue)
* Attaching topic(s) information to analyzed context
* Maximizing the relevancy of the recommended ads

# Project Description

Nowadays, in an information centric society, we are flooded with data (the so called “[deluge of data](http://www.economist.com/node/15579717)”). In this context, an automatic identification of entities that satisfy the user’s information needs is paramount. The goal is to expose to the user only meaningful information (relevant and of interest), more important, at the right time and in the right context [2].

Our project tackles this problem, the so called “context match” between an active context, a possible advertisement and the user that is currently browsing that context.

From the technical point of view, we employ a unified technique for the recommendation of relevant ads. This is why we extend the concept of a “topic” to describe all the three actors from our model. An advertisement is described, ideally, by a *single*, targeted “topic”. The [active context](#currentactive) is modeled by a *static* set of “topics”. Finally, a user is described by a *dynamic* set of “topics” due the evolution over time of interests.

This separation generates the possibility of recommending advertisements from two different perspectives. The first, that maximizes the similarity between the “topics” describing the active context and the advertisement is considered to be a *contextual* recommendation and will influence the page topic coverage. The latter, that considers the best match between the user’s current interests and an advertisement is said to be a *behavioral* recommendation [6] and improves the user interest coverage.

We previously stated that the concept of a “topic” is extended. This generalization is possible due to the fact that, conceptually all the three entities can be mapped on a bag of words model. The advertisement is defined by a “bid phrase” that defines its intended audience. The active context is described by the keywords extracted from its content. Finally, the user is described by his historical information. This history defines a time-variant succession of interests by considering the underlying topics distribution of the visited active context.

The *page topic coverage* and *user interest* *coverage* are antagonistic by nature because they both compete for the same page advertisement slots. This is why we need to balance them by choosing the ones with the maximal relevance in the given overall context.

# Related Work

A common approach in the literature is to describe the matching content with some relevant keywords. These keywords are lexically compared with the descriptors of the ads (bid phrase) hence obtaining a lexical similarity [1, 9]. Such an approach follows a pipeline with a few, well-defined stages. A pre-processing stage is needed to prepare the content by sanitizing, removing stop words, stemming and extracting some keyword candidates (words from the context, annotated with some descriptive features). Then the annotated keyword candidates are ran through a trained classifier. In this way, the keywords are extracted.

Such an approach is generally enhanced with additional models that sustain the idea of a semantic similarity between topics and ads [5, 8, and 10]. This association generates a second score called semantic score that, combined with the lexical, reinforce the match. This semantic information can be embedded in a taxonomy [5] and used to score the similarity based on the distance to the least common ancestor, if both the context and the advertisement is can be mapped on it.

The third aspect to be considered in such a model is the actual user. The associated historical information, if present, will influence the final matching [3, 12]. In general, user information can be attached to the advertisement or to the page [12] but, in recent articles, the idea of user interest and trend is explored [3]. Such a model can extract the dynamics of behavior and make better recommendations.

The concept of a “topic” is described using a specialized [*mixed membership model*](#mmm) called *topic model.* Such a model describes the hidden thematic structure [7] in large collections of documents. The [Latent Dirichlet Allocation (LDA)](#lda)[4] is the simplest of the topic models. The main idea behind LDA is that a document exhibits multiple topics [7]. These topics fully describe the document and represent a distribution over the whole set of available topics (within the document corpus). This distribution was subconsciously modeled by the author of the document, who wanted to transmit a message (collection of words) about an area of interest (the distribution over all topics) using area specific words (distribution of words over topics). All the used words are just sampled from the topic’s word distribution. In this context, the only observable data is the document’s words. The topics, their distribution in documents and the distribution of words between topics need to be inferred. Direct inference is not tractable so approximation techniques are used [11].

# Necessary Resources

Our system has mainly software dependencies:

* We need a low latency, scalable and fast data store for our data-driven computation (chose MongoDB).
* We need a proven relational DBMS for the part of our database schema that has a relational nature (chose MySQL)
* We need an external web NLP API for the initial gathering of candidates for our classifiers – keywords use-case - (chose AlchemyAPI)
* We need fast, low memory footprint HTML Parser (chose JSoup)
* We are using a Java based application development framework (Spring)
* We need a machine learning software for the classifiers (Weka)

We also need an extensive dataset:

* a collection of web pages that will be part of our training/validation/testing set
* a collection of advertisements that will be associated with the context
* a bag of words model that will describe the associated topics from the keywords perspective
* a collection of historical information for a set of identifiable users

NOTE: On this project, I am working with **Octavian Lucian HASNA**, from the Romanian Computer Science line of study.

# Expected Results

We expect to obtain:

* Close to none *false positives* recommendations (in our scenario, an “out of context” ad is worst than no ad at all)
* Maximal topic relevance for the recommendations based on [the given context](#currentactive) and user
  + Increased page topic coverage based on the recommended ads
  + Increased user interest coverage based on the recommended ads
* System scalability (based on the complexity of the online flow)

Due to the subjective nature of the underlying problem and to the lack of annotated benchmark datasets we chose to gather annotated collections of recommendations offered by a commercial system. The performance of the system will be measured based on the distance between the generated recommendations and the ones offered by a commercial system, on the same context. The distance will be computed using the topic model; hence it will represent a distance between two distributions over topics.

# Project Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Week | Period | Research | Implementation | Writing |
| 1 - 8 | 1 nov – 23 dec | Bibliography Study | Comparison & Tests | Weekly short reports |
| 9 - 13 | 3 jan – 3 feb | - (session) - | - (session) - | - (session) - |
| 14 - 15 | 7 feb – 18 feb | Keyword extraction | Integration with external API | - |
| 16 | 20 feb – 29 feb | Keyword extraction | Keyword Classification |  |
| 17 | 1 mar – 7 mar |  | Full Flow Conceptualization | Detailed Component Overview |
| 18 | 8 - 14 | TopicExtraction(TE) | - | TE Methods Overview |
| 19 | 15 –21 | TopicModels-LDA | LDA tests & prototype |  |
| 20 | 22 –28 |  | LDA tests on our model |  |
| 21 | 29 mar - 4apr |  | LDA integration with system |  |
| 22 | 5 - 11 |  | Integrate keywords & LDA | Introduction |
| 23 | 12 – 18 |  | Integrate with ontologies | Bibliographic study |
| 24 | 19 – 25 |  | [Online Flow](#of) Integration | Analysis & design |
| 25 | 26 apr – 2 may |  | [Offline Flow](#offflow) Integration |  |
| 26 | 3 – 9 |  | Implementation stabilization | Implementation |
| 27 | 10 -16 |  | Fine Tuning & Testing |  |
| 28 - 29 | 17 -27 may |  | Performance & scalability | Tests & conclusions |
| 30-31 | 28may-7june | - (session) - | - (session) - | - (session) - |
| 32-33 | 8june-24june |  | Final tests and demo | Improve & Finalize doc |

**Table 1: Overview of the license project timeline until submission**

**NOTE: In the upper table, Wednesdays are considered ends of a week (meeting day with our advisors).**

**NOTE: The *research* column describes the topic/s we investigated during the associated time span. The *implementation* column describes activities involving some actual code. The *writing* column describes some writing effort to summarize / describe the corresponding research / implementation activity + the actual project documentation.**

# Table of Contents

For the table of contents, please see the <FlorinMacicasan_contents_bibliography.pdf> attached document.

# Bibliography

For the bibliography, please see the <FlorinMacicasan_contents_bibliography.pdf> attached document. I referred, in the proposal, the articles by their index in the upper mentioned references.

# Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Current active context | The *web page* with which the user currently interacts. *In extenso*, it describes any meaningful collection of words. |
| Context portion with high descriptive value | We refer to the concept of a *keyword* that outlines the lexical significance within the host document. |
| LDA | Latent Dirichlet Allocation = generative probabilistic model for collections of discrete data |
| Mixed Membership Models | The basic idea is that data is grouped. Each group is modeled with a mixture. The components of the mixture are shared between all the groups. The mixture proportions vary from group to group [14] |
| Online Flow | Describes the real-time use case in which the system maximize the triple similarity relation and offers recommendations |
| Offline Flow | Describes the periodical model calibration (based on newly acquired data) and advertisement / web page topic categorization |

NOTE: This document is based on the proposed template for the 2 page license thesis summary (http://cs.utcluj.ro/csd/pics/diploma/templateDoc-Eng2011.zip).