System for building and analyzing preference models based on social networking data and SAT solvers

Radosław Klimek

AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
rklimek@agh.edu.pl

Abstract. Discovering and modeling preferences has an important meaning in the modern IT systems, also in the intelligent and multi-agent systems which are context sensitive and should be proactive. The preference modelling enables understanding the needs of objects working within intelligent spaces, in an intelligent city. There was presented a proposal for a system, which, based on logical reasoning and using advanced SAT solvers, is able to analyze data from social networks for preference determination in relation to its own presented offers from different domains. The basic algorithms of the system were presented as well as the validation of practical application.

Keywords: preference model; SAT solvers; social networking data; Facebook, Twitter

1 Introduction

Modelling users' preferences enables to support decision making. The choice of decision can be carried on in an interactive process which describes user's aspirations and goals. Those aspects can have a fundamental meaning in intelligent systems which need to work pro-actively, understand their own context and undertake actions which will improve comfort and safety of city residents. Also the companies try to get data to build profiles of their present and potential clients. The building process of such profiles can be automatized which helps to avoid costly and time-consuming market surveys.

The priceless source of information are social networking platforms and data in form of posts, photos, statements, opinions, circles of friends, etc. This type of information is quite easily accessible and hardly removable.

The aim of this work is to propose a system which builds online and in real time preference models. Later on, those models can be adjusted to a particular offer by logical reasoning based on accessible SAT solvers. The project and system can find a wider use in relation to multi agent systems where preference models are built on the basis of results which have already been required. The next stage is choosing, in logical reasoning process, the agents which meet best certain expectations. All of the processes are carried on fully automatically.

Recommendation system are always at the center of interest, see [5], where a preference model are build to find the user neighbor set. Also, mining online reviews and tweets for predicting [6] have an impact on the future sales. Preference models support sensing and decision-making behavior analysis [1]. However, there is a lack of works with preference models involving social media mining and SAT-based analysis.

2 Functional model

At the beginning the functional model of the system will be presented, see Figure 1. Some used terms will be discussed in the next section. The following actors

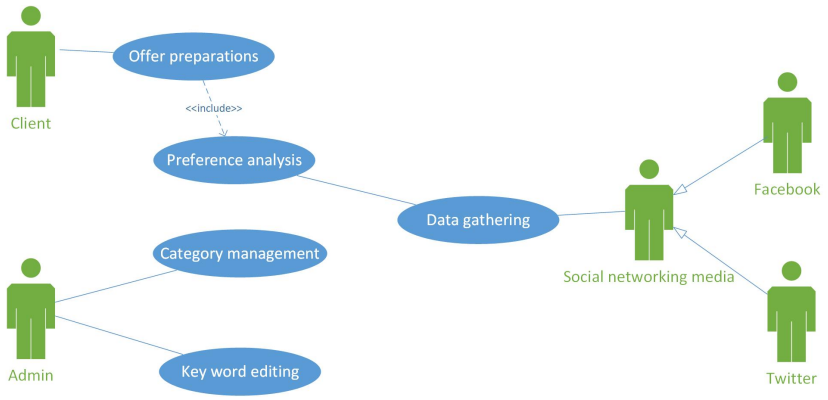

Fig. 1. The use case diagram for the proposed system

are taken into consideration:

- **Client** – a person or another system sending request to the system, the goal is to get an offer for presented preferences. It includes a possibility of giving authorization data which enables logging and getting an access to data on social networks which are the sources of basic (mass) data.
- **Admin** – a person who is empowered to modify knowledge database, categories and particular key words, adding the new and editing the already existing ones, deleting outdated records or incorrect entries.
- **Social networking media** – networking site which gathers and shares user's data. It is a source of basic pieces of data which is furtherly processed and the preferences are formulated on the basis of logical reasoning.

The following use cases are offered:

- **Offer preparations** – preparation of an offer from a particular branch, for example: tourism, real estate market, etc. This data is transformed into a version which enables logical reasoning.

- **Preference analysis** – the process of logical reasoning which aim is to compare data collected from different social networks and data from prepared offers. Among the searched positions are those which have the maximum of compliance in relation to other offers and already collected data.
- **Data gathering** – Data gathering – collecting the basic data from particular social networks on the basis of well-known and widely accessible searching algorithms, mentioned in Section 3.
- **Key word editing** – editing single key words, adding the new ones, deleting outdated words and modifying them.
- **Category management** – Category management – the process of managing the categories which enable sorting the key words.

The following scenarios are build, see Tables 1–5:

Table 1. Scenario: adjusting the offer to the user's preferences

Adjusting the offer to the user's preferences:
Trigger event: a client sends a request to the system. Preconditions: the system should have a necessary amount of offers and key words in its database.
Scenario: 1. Client sends a request to the system together with giving a token. 2. «include» Use case: searching for user's preferences. 3. Creating a logical formula on the basis of user's preferences and the data from an offer. 4. Finding formula quantifying. 5. The system sends back preferences and a proposed offer.
Postconditions for success: in a request there should be sent active tokens of the user Postconditions for failure: disabled user tokens, too small amount of data in social media, insufficient database of the own offers.

3 Data and preferences modelling

The draft version of interests and preferences map for a particular networking site is built on the basis of existing and well-known algorithms which will be discussed here very briefly.

One of those algorithms is *Latent Dirichlet Allocation* LDA intended to process natural language [2]. It is a generative statistical model. On the input all text documents are analyzed. On the output we get topics with key words which fit best into the context of analyzed documents. The algorithm treats every document as a collection of different topics. The topics are described by a set of key words. LDA algorithm was used to determine the topic of documents from the year 2003 by David Blei, Andrew Ng and Michael I. Jordan.

Table 2. Scenario: searching for user's preferences

Searching for user's preferences:
Trigger event: a client sends a request to the system. Actors: a client, social media (for example Facebook, Twitter) Shareholders and their goals: a client wants to get user's preferences. Social media share their data. Preconditions: the system should have a necessary amount of offers and key words in its database.
Scenario: 1. Client sends a request to the system together with giving a token. 2. «include» Use case: downloading user's data 3. Searching for user's preferences. 4. System sends back preferences on the basis of well-known algorithms.
Postconditions for success: in a parameter there should be sent an active token of the user. Postconditions for failure: disabled user tokens, too small amount of data in social media, insufficient amount of key words.

Table 3. Scenario: downloading user's data

Downloading user's data:
Actors: social media (Facebook, Twitter) Shareholders and their goals: social media (Facebook and Twitter) share their data. Trigger event: system sends a proper request. Preconditions: possibility of establishing a connection with social media
Scenario: 1. System sends a request with an active user token. 2. Social media share their data through their API.
Postconditions for success: in a request parameter there should be sent an active token of the user. Postconditions for failure: disabled user token, too many requests in a limited time period.

Another approach is presented by algorithm *Bag-of-words model* BOW – simplified representation of documents in processing natural languages and looking for information). It enables to easily classify documents. In BOW model text is written as a set (bag) of words without paying attention to their grammar and sequence. The important factor is amount repetitions of particular words. The BOW model was described in an article by Harris [3].

This kind of approach in both algorithms is comfortable and proper for our designed system. We assume thematic division on categories and key words. Categories match topics. Key words match concepts. However, it is possible that one key word can be attributed to more than one category. The example is "swimming" which can be attributed to „sport" category but also to "seaside holiday"

Table 4. Scenario: adding a category

Adding a category:
Actors: Administrator Shareholders and their goals: the aim of administrator is to add a new preference category Preconditions: a person has administrator privileges.
Scenario: 1. System displays a formula on its website 2. Administrator fills in the form and sends it back. 3. System confirms adding a new category.
Postconditions for success: the particular category does not exist in a system. Postconditions for failure: a category has already existed in the system.

Table 5. Scenario: Adding a key word

Adding a key word
Actors: Administrator. Shareholders and their goals: the aim of administrator is to add a new key word to a category. Trigger event: Administrator sends a formula to the system. Preconditions: a person has administrator privileges.
Scenario: 1. System displays a formula on its website. 2. Administrator fills in the form and sends it back. 3. System confirms adding a new key word.
Postconditions for success: the particular key word does not exist in a system. A category is attributed to the key word. Postconditions for failure: a key word has already existed in the system.

category. The exemplary categories, analyzed in this work, were presented in Table 6. The following exemplary offers are presented in Tables 7-9.

Data acquisition from social media requires API knowledge. In case of Facebook it is REST endpoint and it is called Graph API. In this way we can get descriptions of websites liked by the user. Authentication purposes require generating a token with the user's consent. The exemplary listing with a request was presented in Listing 1.1. As the answer we can get tables of objects which represent particular websites. Each of those objects has website id and its description. A collection of descriptions of all pages liked by the user creates input data for algorithm looking for preferences. The exemplary answer was presented in Listing 1.2.

The similar procedure can be repeated in case of Twitter. According to API documentation a request is sent together with Ouath authentication heading which was presented in Listing 1.3. The answer contains of table of objects expressed in JSON format. The important fields are tweet contents, see Listing 1.4. Those pieces of data are an input for programs mentioned above which gather data about preferences.

Listing 1.1. Request for favorite pages (Facebook)

```
https://graph.facebook.com/v2.8/me?
access_token=EAAKldGZCnym4BAKCPpkSjiDHvv2NSU6jWmOgpZCmPRH
0vQloryvuZBQ8Jrb5uliBZARkheALC8fmHGuCTMclnbbaw82AfLotKFRc
kXZCqv4NKEMKMPEvl33NvyuFwGtoXmZAVIG4LDZBgNuiB9YsZCGxrw5aZ
CjaFw6XZA7Q7eF6TZA8lzpgZCB7rnkaZA9R9KaBRjp8ZD
&callback=FB.__globalCallbacks.f2ced5f8a183a7c
&fields=likes%7Babout%7D &method=get
&pretty=0 &sdk=joey
```

Listing 1.2. Answer with favorite pages (Facebook)

```
{"likes": {
 "data": [ {
   "about": "Java and JVM based technologies, dynamic
          languages, RIA, enterprise architectures,
          patterns, distributed computing and much more...",
   "id": "354953985700"
          },
  { "about": "Interesting Engineering is a cutting edge,
           leading community designed for all lovers of
              engineering, technology and science.",
    "id": "139188202817559"
          },
  { "about": "Pasja. Każdy ma swojąc?\n\n Prywatna
        strona o szeroko pojętych nowych technologiach",
    "id": "125612537464918"
        },
  {"about": "Koło Naukowe BIT gromadzi najlepszych
      studentów Informatyki na Akademii Górniczo-Hutniczej
      w Krakowie. Odwiedź naszą stronę aby dowiedzieć się więcej!",
   "id": "719620231484798"
       },
  {"about": "Sekcja Koła Naukowego BIT, która tworzy framework
      do zdobywania wiedzy i realizacji własnych pomysłów. ",
   "id": "1535988496663169"
      },
  {"about": "This page is managed by Marcin Marczewski and
      Jakub Hankiewicz. It follows the IBM Social
          Computing Guidelines.\n ",
   "id": "348362861915681"
    } ] } }
```

Listing 1.3. Request for favorite tweets (Twitter)

```
https://api.twitter.com/1.1/favorites/list.json?count=2&
screen_name=vapsel21&include_entities=false}
```

Table 6. Cattegories and key words

Category	Key word
Sport	cycling, football, skiing, basketball, volleyball, box, running, hockey, swimming, canoeing, climbing.
Drinks	coffee, tea, juice, beer, wine, water, cocktail.
Food	bread, apple, banana, beef, pork, veal, bacon, borscht, black pudding, French fries, cookies, duck, fish, vegetarian.
Nature	sea, mountains, forest, river, lake.
Dwelling	flat, hotel, hostel, apartments, shelter.
Premises	cafeteria, restaurant, bar, pub, fast food restaurant, inn.
Transport	plane, car, bike, bus, helicopter, motorbike, scooter, ship, walking.
Entertainment	cinema, bowling, billiard, snooker, theater, dance, museum, zoo.
Season	spring, summer, autumn, winter.
Travelling	camping, snorkeling, yacht, climbing, sights.
Facilities	air conditioning, phone, Internet, Wi-Fi, computer.

Table 7. Exemplary tourist offer: a trip to Morskie Oko

Category	Parameters
Sport	climbing
Drinks	coffee
Nature	mountains, forest, lake
Dwelling	shelter
Transport	car, walking
Season	spring, summer, autumn, winter
Travelling	camping, sights

Table 8. Exemplary tourist offer: Beach volleyball tournament

Category	Parameters
Sport	volleyball
Dwelling	hotel
Nature	sea
Transport	bike, walking
Season	summer

Table 9. Exemplary tourist offer: Ski station

Category	Parameters
Sport	skis, hockey
Drinks	tea, coffee
Nature	mountains, forest
Dwelling	hotel, apartments.
Premises	inn, restaurant.
Transport	car, bus, walking.
Season	winter

Listing 1.4. Answer with favorite tweets (Twitter)

```
[
{"created_at": "Wed May 17 08:56:03 +0000 2017",
"id": 864766453119152129, "id_str": "864766453119152129",
"text": "@GeeCON starts today in Krakow! Participants were
welcomed by David Moore, SVP of TN Product Development.
https://t.co/ CGTCoEpiJJ", ...
},
{"created_at": "Tue Apr 11 22:52:43 +0000 2017",
"id": 851931044789886976, "id_str": "851931044789886976",
"text": "See all sessions from this year's ng-conf 2017
with Angular core team and many many others from around
the world. ... https:// t.co/s43U6O2yir", ...
}
]
```

The basic processing algorithm, in relation to data gathering, can be presented in a following way:

1. The user agrees to download data from social networks.
2. Collecting data from social networks.
3. Processing collected data according to BOW model.
4. Filtering out short words, links and emoticons.
5. Looking for key words from glossary.
6. Determining the frequency of words. When they exceed threshold limit, they are marked as preferences.

The users permission is connected with sending authorization data. In case of Facebook, the subject of analysis are liked websites and in case of Twitter – published posts. The data is analyzed on the basis of programs, mentioned above, which look for key words and topics. The texts are divided into single words. Furtherly, the categories and key words are determined. The threshold limit, based on word frequency and necessary to determine preferences, can be changed at any time. In current experiments it was set as number 3.

4 Offer analysis with the use of solvers

In order to perform preference analysis, together with assessment of notified offers, SAT solvers will be used. The SAT satisfiability problem is a classical IT problem which challenge is to find a substitution satisfying certain logical formula. For propositional calculus it will be substitution for sentential variables. Nowadays, the big development in searching the whole space of states was made and it routinely solves tasks consisting of tens of thousands of variables. There exist ready-to-use SAT solvers. The problem similar to SAT is MaxSAT problem based on finding a substitution which will maximally satisfy a formula. Thus, it considers the possibility of nonexistence of (full) satisfying substitution but in such case we look for substitution for the biggest possible number of clauses.

The system will have embedded MaxSAT solver in a form of SAT4J library [4]. The formulas will be saved in CNF form. The exemplary offer, presented in Table 6 written as a logical formula, will be as follows:

$$\begin{aligned} O = \neg cycling \wedge \neg football \wedge \neg skiing \wedge \neg basketball \wedge \neg volleyball \wedge \neg box \\ \wedge \neg running \wedge \neg hockey \wedge \neg swimming \wedge \neg canoeing \wedge climbing \wedge coffee \\ \wedge \neg tea \wedge \neg juice \wedge \neg beer \wedge \neg wine \wedge \neg water \wedge \neg coctail \wedge \neg bread \\ \wedge \neg apple \wedge \neg banana \wedge \neg beef \wedge \neg pork \wedge \neg veal \wedge \neg bacon \wedge \neg borscht \\ \wedge \neg blackpudding \wedge \neg Frenchfries \wedge \neg cookies \wedge \neg duck \wedge \neg fish \\ \wedge \neg vegetarian \wedge \neg sea \wedge mountains \wedge forest \wedge \neg river \wedge lake \\ \wedge \neg dwelling \wedge \neg hotel \wedge \neg hostel \wedge \neg apartments \wedge shelter \wedge \neg cafeteria \\ \wedge \neg restaurant \wedge \neg bar \wedge \neg pub \wedge \neg fastfoodrestaurant \wedge \neg inn \\ \wedge \neg plane \wedge car \wedge \neg bike \wedge \neg bus \wedge \neg helicopter \wedge \neg motorbike \wedge \neg scooter \\ \wedge \neg ship \wedge walking \wedge \neg cinema \wedge \neg bowling \wedge \neg billiard \wedge \neg snooker \\ \wedge \neg theater \wedge \neg dance \wedge \neg museum \wedge \neg zoo \wedge spring \wedge summer \\ \wedge autumn \wedge winter \wedge camping \\ \wedge \neg snorkeling \wedge \neg yacht \wedge \neg climbing \wedge sights \wedge \neg airconditioning \\ \wedge \neg phone \wedge \neg Internet \wedge \neg Wi - Fi \wedge \neg computer \end{aligned} \tag{1}$$

Suppose that we have preferences for a user expressed by a formula:

$$P1 = skiing \wedge coffee \wedge mountain \wedge walking \wedge winter \tag{2}$$

After converting the offer as well as users preferences to logical formula, we can write the final formula which can be used as an input to MaxSAT solver and has a form as follows:

$$O \wedge P$$

If MaxSAT solver finds a proper assignements to satisfy this formula, it means that it meets all users preferences and can be recommended. This algorithm can be repeated involving all offers from database for every user.

Suppose that in the base we have three offers, as in Tables 7–9. For a user described by Formula 2 the system will recommend „Ski station" option because it meets all user's preferences. The offer "Trip to Morskie Oko" does not satisfy „skiing" parameter and the offer „Beach Volleyball Tournament" – skiing, coffee, mountains and winter. For another user described by a formula:

$$swimming \wedge volleyball \wedge juice \wedge sea \wedge river \wedge summer \wedge spring \wedge diving$$

the system will not be able to propose an offer because none of them satisfy user's preferences.

The proposed system will be REST server receiving data necessary to integrate with external servers and algorithm working results. The input data for server is presented by RequestParametersDTO class (Figure 2). The system will download users data by FacebookGraphAPIv2.9 and Twitter API1.1. After receiving a request, server downloads user's data from a social network and sends it for preference searching and later on to MaxSAT solver module. A class with server response model is presented in Figure 2. The first answer field includes a name of proposed offer. The second one is a list of user preferences sorted according to probability of existing of such preference.

The system will consist of a few modules:

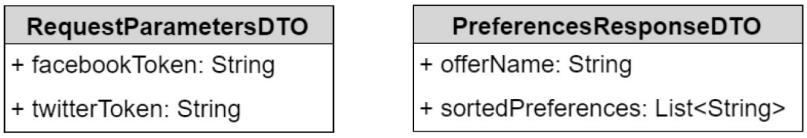

Fig. 2. Class RequestParametersDTO and its parameters for the server. Class PreferencesResponseDTO and its response model for server

- Core – module responsible for handling the operations connected with database. In a project PostgreSQL 9.4.10 is used. Database is installed on the same machine as server with web module. In this module there is also implemented algorithm of searching user's preferences based on ideas of LDA algorithms.
- MaxSAT solver – module used to convert offers and preferences to logical formulas, later on MaxSAT solver (treated as a "black box") is launched.
- Web-module: communicates with clients and social networks by REST service.

5 Conclusions

In this paper we propose some solutions that refer to preference modeling of social network data on the basis of logical reasoning using available solvers. This solution is very efficient because modern solvers are able to solve very quickly tasks consisting of hundreds of thousands of variables. We can imagine a system which processes the flows of data online and in real time. It can be used in different branches (tourism, real estate market, etc.) and in different use classes such as multi agent systems competing for resources.

The further works should concentrate on building a prototype version of the system, another algorithms of building preferences (namely data acquisition form social networks and its analysis), including new logical reasoning schemes (deductive reasoning, modus ponens rule and the others) and another social networks. The opportunities for research and implementation of the proposed system are really wide and promising.

References

1. Abdar, M., Yen, N.Y.: Design of a universal user model for dynamic crowd preference sensing and decision-making behavior analysis. IEEE Access 5, 24842–24852 (2017)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
3. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
4. Le Berre, D., Parrain, A.: Website: Sat4j – the boolean satisfaction and optimization library in Java (2017), http://www.sat4j.org/, accessed on 8-Jun-2017
5. Liu, Y., Xie, Q., Xiong, F.: Recommendations based on collaborative filtering by tag weights. In: 2017 13th International Conference on Semantics, Knowledge and Grids (SKG). pp. 62–68 (Aug 2017)
6. Magdum, S.S., Megha, J.V.: Mining online reviews and tweets for predicting sales performance and success of movies. In: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 334–339 (June 2017)