Primena istraživanja podataka za otkrivanje obrazaca u oblasti kulinarskih recepata

Nevena Nikolić
student akademskih master studija
Fakultet tehničkih nauka
Novi Sad, Srbija
nevena.nikolic.ns@gmail.com

Bernadeta Ralbovski
student akademskih master studija
Fakultet tehničkih nauka
Novi Sad, Srbija
bernadeta_ralbovsky@hotmail.com

Kristina Pejić
student akademskih master studija
Fakultet tehničkih nauka
Novi Sad, Srbija
christina_pejic@hotmail.com

Sadržaj— **Svedoci smo procvata broja servisa na Internetu koji korisnicima nude različite recepte za pripremu hrane i pića. Sistemi koji bi omogućili automatsku analizu recepata, kao i detekciju obrazaca i zakonitosti koji se u ovom specifičnom skupu podataka javljaju imaju značajan potencijal za primene i mogućnost unapređenja ovih servisa. U radu je opisana primena algoritama za otkrivanje pravila asocijacije kako bi se, otkrili obrasci koji se najčešće javljaju u kulinarskim receptima. Prikazana studija bazirana je na javno dostupnim receptima za pripremu piva. Upotrebom istraživanja podataka, utvrđeni su najčešći načini na koji se određeni sastojci koriste. Izdvajanjem i prikazom zakonitosti, može se generisati i novi, „univerzalni" recept za pivo, predstavljen u vidu optimizovane mešavine svih vrsta piva i njihovih načina pripreme.**

Ključne reči - Data Mining; pravila asocijacije; Apriori algoritam

I. UVOD

Svakim danom sve više raste popularnost veb sajtova na kojima se pronalaze kulinarski recepti. Kulinarski recepti, danas, preplavljuju Internet u različitim formama, uključujući tzv. kuvare, kuvarske enciklopedije i on-lajn igrice. [1] Na raznovrsnim sajtovima pronalaze se kompleksne baze podataka, bogate idejama i receptima, pružajući mogućnost identifikacije najfrekventnijih sastojaka datih recepata.

Primena tehnika istraživanja podataka na podatke javno dostupne putem ovih servisa omogućava pružanje dodatnih usluga u smislu automatske identifikacije najreprezentativnijih recepata i generisanja novih recepata. Cilj studije opisane u ovom radu je upravo razvoj pristupa pronalaska univerzalnog recepta. Pristup je zasnovan na primenu algoritma za pronalaženje pravila asociranja podataka nad bazom podataka prikupljenih sa javnih Internet servisa, čija je tematika pravljenje piva. Cilj rada je otkrivanje inicijalnog recepta od kojeg su ostali derivirali. Za polazni, univerzalni recept je smatran onaj koji sadrži najveći broj karakteristika zajedničkih svim receptima. Analiza podataka je sprovedena uz pomoć programskog rešenja Weka.

II. PREGLED RELEVANTNE LITERATURE

Priprema hrane i pića spada u kompleksnu aktivnost, koja je bazirana na znanju, stečenom u toku samog procesa pripreme. Mnogo je istraživanja i radova koji se bave upravo ovom temom, oslanjajući se na različite metode [2][3]. Primena informacionih tehnologija i veštačke inteligencije u domenu kulinarstva je ipak ograničena. Hashimoto i saradnici [4] opisuju sistem za podršku kuvanju, zasnovan na prepoznavanju operacija pripreme hrane uz pomoć kamere. Na osnovu ove informacije razvijen je sistem koji u odgovarajućim vremenskim razmacima kuvaru prezentuje informacije vezane za operaciju koja sledi.

Keisuke i saradnici su predložili koncept automatskog stvaranja video sekvenci za podučavanje kuvara na osnovu prepoznavanja operacija kuvanja u tekstualnom i multimedijalnom sadržaju [5].

Istraživanje podataka se često koristi za unapređenje sistema pretrage recepata [6], kao i u sistemima za preporučivanje novih recepata korisnicima [7].

Neka istraživanja su odvela toliko daleko, da se projektovao sistem, koji, u zavisnosti od posluženog menija, generiše boje i dizajn kuhinjskog stola[8].

Primena istraživanja podataka za otkrivanje univerzalnog recepta nije, koliko je nama poznato, opisana u literaturi.

III. OPIS METODA RADA

Apriori algoritam predstavlja klasičan, vrlo koristan algoritam za otkrivanje pravila asocijacije [9], koji je korišćen u ovoj studiji.

U praksi, podaci su uglavnom sirovi i verovatnoća pronalaženja pravila u takvom skupu je veoma niska. Istraživanje pravila može da se izvede nekoliko puta, svaki put sa različitim parametrima, kako bi se povećala verovatnoća nalaska odgovarajućih, netrivijalnih pravila. To je domen, gde je jednostavnost i lakoća generisanja ogromnog skupa pravila Apriori algoritma, gotovo nepobediva. Potraga za specifičnim pravilima, ne treba uvek da ide u dubinu, iz tog razloga, što bi daljim specifikovanjem samo pravilo izgubilo na značaju i postalo trivijalno [10].

A. Prikupljanje podataka

Veb sajt, sa kojeg je prikupljena kompletna receptura je www.brew-monkey.com, sadrži recepte struktuirane u formatu XML (*Extensible Markup Language*). Za parsiranje XML dokumenata korišćen je PHP (*Hypertext Preprocessor*) i SimpleXML. SimpleXML je jednostavna, ali vrlo pogodna klasa, koja ima mogućnost kreiranja, čitanja, modifikacije i snimanja XML dokumenata. Funkcioniše tako što kreira SimpleXML objekat iz nekog izvora (string ili fajl, u ovom slučaju fajl). Baza podata, u sebi, ne sadrži nikakav tip normalizacije, kako bi bila pogodna za rad saWeka (*Waikato Environment for Knowledge Analysis*) alatom.

Nakon toga, sledi etapa u kojoj su podaci, smešteni u bazu podataka, konvertovani u CSV (*Comma Separated Value*), a potom u ARFF(*Atribute-Relation File format*)format.

B. Weka

Weka [11] je popularan alat mašinskog učenja, napisan u Java programskom jeziku. Weka sadrži kolekciju alata za vizuelizaciju i algoritme za analizu podataka i modelovanje na osnovu predviđanja, uključujući grafički interfejs, što u velikoj meri proširuje mogućnosti korišćenja, odnosno, broj mogućih korisnika. Podržava većinu zadataka istraživanja podataka uključujući prvobitnu obradu podataka (eng. *preprocessing*) i otkrivanje pravila (eng. *Association Mining*). ARFF fajl predstavlja tekstualni format koji Weka koristi za čuvanje podataka u bazi podataka.

Glavni interfejs jeste, grafički, a za potrebe ovog rada korišćen je pristup pomoću komandne linije, dok su zadate komande preuzete iz grafičkog interfejsa.

ARFF fajl predstavlja tekstualni format koji Weka koristi za čuvanje podataka u bazi podataka.

C. Apriori algoritam

Detaljnim poređenjem različitih algoritama pogodnih za pronalaženje asocijacija i pravila (*Apriori,* ECLAT*, FP-growth...*) ustanovljeno je da je za potrebe ovog rada najpogodniji Apriori algoritam, iz više razloga, a pre svega zbog toga što daje odgovor na pitanje – koliko često se pojavljuju instance atributa u bazi podataka.

Apriori algoritam predstavlja klasičan algoritam za pronalaženje čestih skupova podataka i iznalaženje pravila asocijacije nad transakcijonom bazom podataka. Radi na osnovu identifikovanja frekventnih podataka u bazi, proširujući ih u sve veće skupove. Skup podataka, određen Apriori algoritmom, koristi se u pronalaženju pravila asocijacije, s akcentom na trendove u bazi podataka [12].

Istraživanje pravila asocijacije pronalazi interesantne veze ili korelacije nad širokim skupom podataka [13]. Formalnija definicija je:

Neka je $I=\{i_1,i_2, \ldots, i_m\}$ skup podataka i D, skup transakcija baze gde je svaka transakcija T skup koji zadovoljava $T \subseteq I$. Asocijativno pravilo je onda $A \Rightarrow B$, gde je $A \subset I$, $B \subset I$ i $A \cap B = \emptyset$.

1) Podrška (eng. Support)

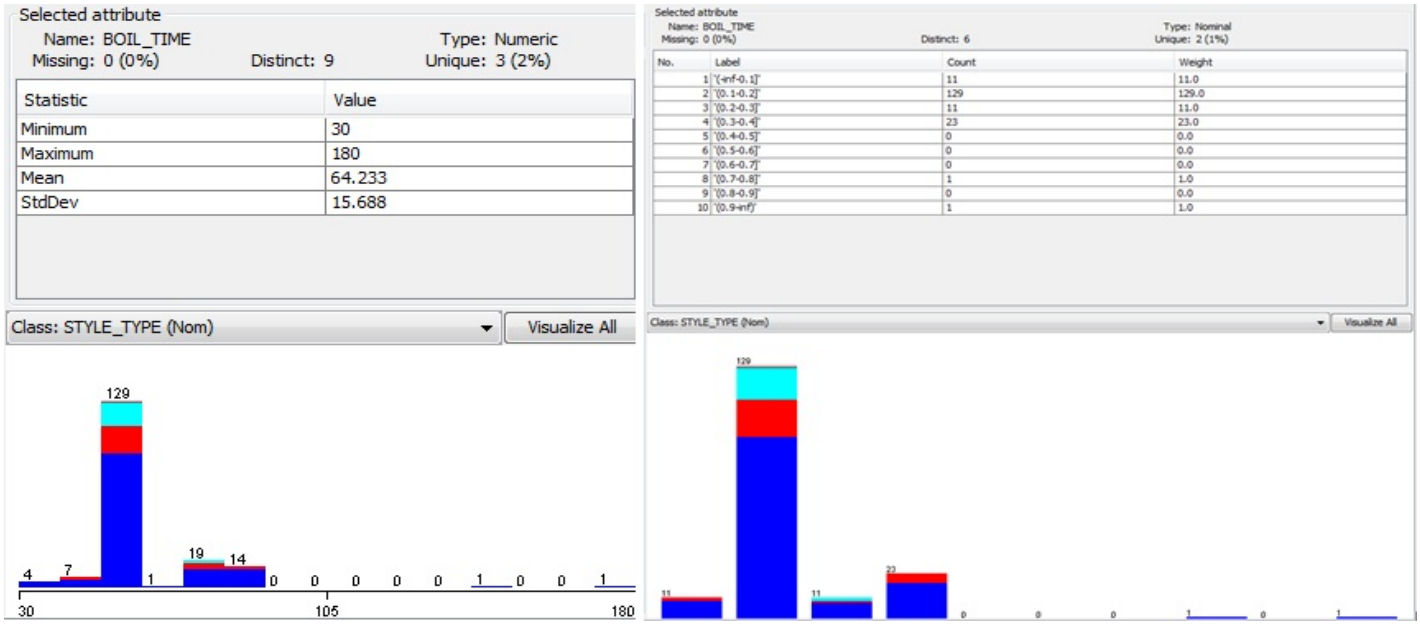
Pravilo $A \Rightarrow B$ u transakcionom skupu D sadrži podršku s, gde je s procenat transakcija u transakcionom skupu D, koje sadrže $A \cup B$ ili i A i B. Nadalje, ovo postaje verovatnoća $P(A \cup B)$ [8].

1) Poverenje (eng. Confidence)

Poverenje pravila $A \Rightarrow B$ u skupu transakcija D je c, ako je c procenat transakcija koje ako sadrže A, sadrže i B. Ovo predstavlja uslovnu verovatnocu $P(B|A)$.

Pronađena pravila smatraju se interesantnim ukoliko zadovoljavaju minimalnu zadatu vrednost podrške i minimalnu zadatu vrednost poverenja [14].

U konkretnom zadatku za mere *Confidence* i *Support* usvojene su visoke vrednosti kako bi se izvukla pravila koja zadovoljavaju najveći broj pojava u bazi recepata.



Slika 1: Atribut 'Boil time' pre i nakon normalizacije i diskretizacije

IV. REZULTATI RADA

Kako bi sistem mogao biti korišćen više puta i za različite skupove podataka, kreiran je skript fajl sa komandama koje će biti izvršavane uz pomoć komandne linije. Kao ulazni parametri uzimaju se podaci nad kojima se vrši dalje istraživanje u ARFF formatu. Ulazni podaci, vrlo često, mogu biti različitog tipa, odnosno, nestandardizovani za pronalaženje pravila asocijacija među njima putem Apriori algoritma. U fajlu sa podacima, na koje će biti primenjen Apriori algoritam, ne sme da se nađe niti jedna druga vrsta atributa, osim nominalnih.

A. Obrada podataka (eng. Preprocessing)

Weka poseduje filtere, putem kojih je omogućena konverzija atributa. Podaci sa kojima je rukovano, bili su delom nominalni, a delom numerički. Fajl sa podacima propušten je kroz filtere za normalizaciju i diskretizaciju.

1) Normalizacija

Pre puštanja skupa podataka kroz filter za diskretizaciju, urađena je normalizacija. Kao rezultat dobijen je format podataka u kojem se sve vrednosti atributa nalaze u optimalnim (identičnim) opsezima.

2) Diskretizacija

Discretize filter diskretizuje numeričke atribute u nominalne. Na Sl. 1 prikazana je jedan od atributa na osnovu kojih se razlikuju načini pripreme piva tokom pripreme podataka. Sa leve strane vide se sirovi, numerički podaci, a sa desne strane, obrađeni, nominalni podaci, spremni za rad sa Apriori algoritmom.

B. Otkrivanje pravila asocijacija(eng. Association Mining)

Nakon obrade i neophodnog filtriranja podataka, naredni korak jeste komanda vezana za definisanje parametara Apriori algoritma i specificiranje fajla (u daljem tekstu output.*txt*) gde će biti sačuvana rezultirajuća pravila pronađena uz pomoć algoritma. U tekstualnom output fajlu nalaze se najbolja pravila koja su pronađena na osnovu zadatih parametara.

Dobijeno je dvadesetak pravila, od kojih će biti prikazano pet najupečatljivijih, koji su najviše pripomogli u zadatku pronalaženja univerzalnog recepta.

Pravila su sledeća :

1. HOP_USE=Boil 167 ==> FERMENTABLE_ADD_AFTER_BOIL=FALSE 167
2. FERMENTABLE_ADD_AFTER_BOIL=FALSE 175 ==> YEAST_FORM=Liquid 145
3. YEAST_TYPE=Ale 134 ==> YEAST_FORM=Liquid 113
4. HOP_FORM=Pellet 144 ==> HOP_USE=Boil 140
5. FERMENTABLE_TYPE=Grain 105 ==> FERMENTABLE_ADD_AFTER_BOIL=FALSE 105

Ova pravila su tipa ‘ako-onda’. U prikazanim relacijama među atributima i njihovim pojavama na očigledan način uočavaju se zakonitosti vezane za način pripreme piva.

Prvo pravilo odnosi se na upotrebu hmelja i dodavanje aditiva za fermentaciju. Iz prikazanog rezultata, zaključuje se da se u 167 od 176 instanci hmelj prvo kuvao, a aditivi su uvek dodavani pre kuvanja.

U drugom pravilu očigledno je da je u 145 instanci korišćen kvasac u tečnom stanju, kao i to da se aditivi dodaju pre kuvanja, slično prvom pravilu.

Treće pravilo govori da se u 134 instance pojavljuje tip kvasca 'Ale', i slično drugom pravilu, da se dodaje uvek u tečnom stanju.

Četvrto pravilo jasno pokazuje da je u većini instanci, 144 od 176, hmelj korišćen u formi peleta i da se koristi kuvan.

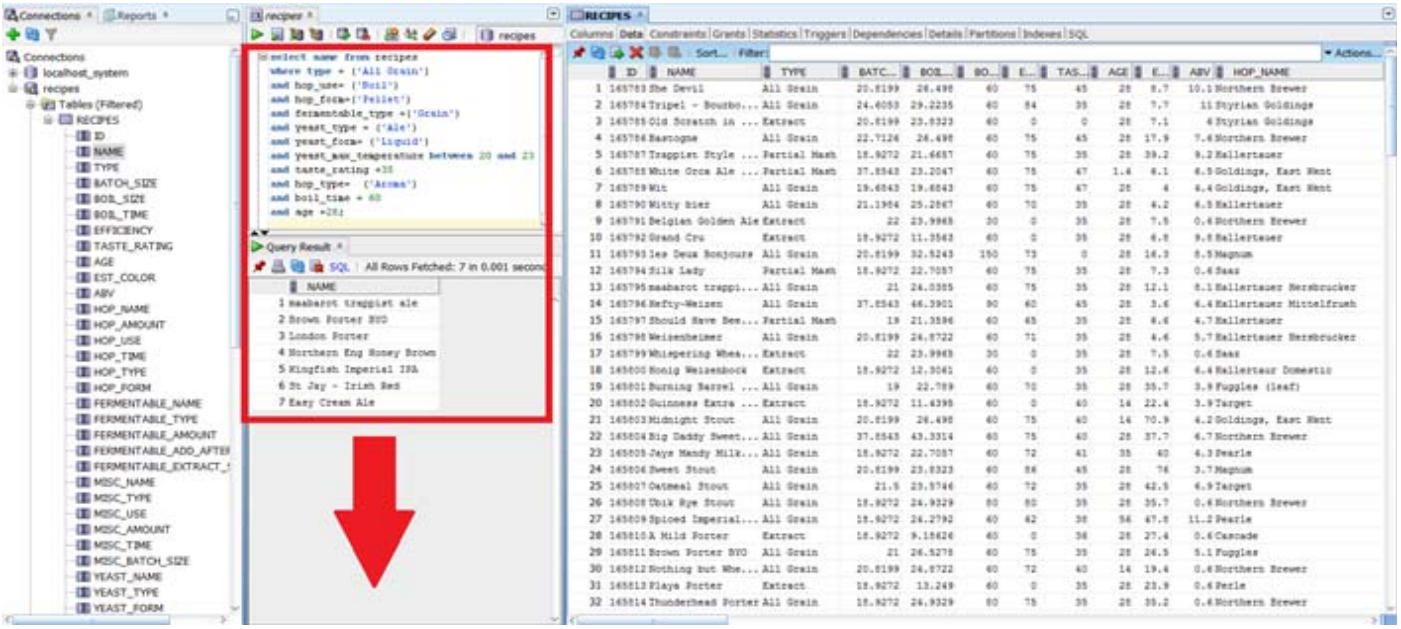
Peto pravilo implicira da se dodaci većinom dodaju u zrnu i, opet, da se aditivi za fermentaciju uvek dodaju pre procesa kuvanja.

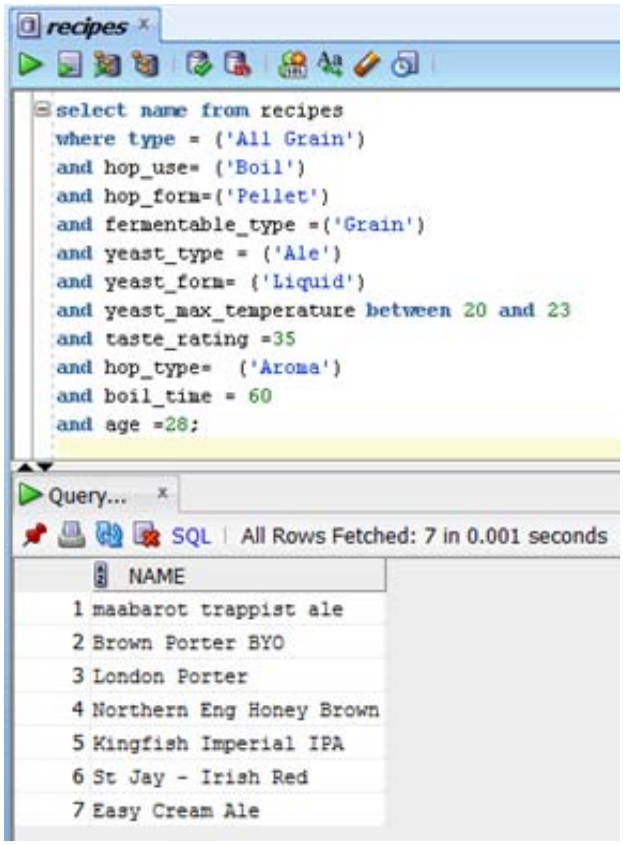
Na osnovu ovih međuzavisnosti i cilja studentskog projekta, dobijena pravila su raščlanjena na takav način da se može strukturirati univerzalnirecept. To bi značilo da bi mogući recept univerzalnog piva mogao glasiti :

- Svi fermentacioni aditivi (žitarice) dodaju se uvek pre kuvanjau neobrađenom obliku(zrno)
- Ubaciti hmelj u peletiranom obliku i kuvati
- Kvasac 'Ale' dodati u tečnom stanju
- Vreme kuvanja je 60-75 minuta i sl.

U prethodnom receptu, prikazani su samo neki od mogućih sastojaka i načina pripreme.

Dvadeset dobijenih pravila iskorišćena su u *select* upitu nad bazom podataka (Slika 2) koja sadrži recepte kako bi se dobio onaj recept koji bi najviše odgovarao univerzalnom. Univerzalni recept dobijen primenom Apriori algoritma zapravo predstavlja način pravljenja osnovnog recepta od kojeg ostali nastali. Kao rezultat, *select* naredbom je dobijeno sedam recepata.





V. ZAKLJUČAK

Istraživanje asocijativnih pravila proizvodi ogromnu količinu pravila, od kojih je većina redundantna. U ovom radu je prikazan jedna moguća primena istraživanja asocijativnih pravila uz pomoć softverskog alata Weka i u nju implementiranog Apriori algoritma.

Odabrani metod primenjen je na 176 instanci, koje opisuju sastojke koji se koriste prilikom pravljenja piva. U idealnom slučaju bilo bi posmatrano više instanci kako bi recept bio što verodostojniji.

Propuštanjem instanci kroz Apriori algoritam, dobijena su pravila, koja su potom pregledana i od kojih su samo najznačajnija izabrana za prikaz i kreiranje *meta* recepta, kako bi se, pre svega, smanjila redundantnost. Pravila su raščlanjena na vrednosti karakteristika i kao takva propuštena kroz bazu kako bi se isfiltrirali recepti i dobili oni koji zadovoljavaju kriterijume.

Ovo rešenje je primenjivo ne samo na pronalazak originalnog recepta, tkz. univerzalnog, za pripremu piva, već i za sve ostale kulinarske recepte. Područje primene rešenja obuhvata sve ono što pojavljuje u više sličnih oblika, a želi se pronaći inicijalna verzija. Inicijalna verzija dobijena primenom rešenja nikako ne znači da je sve što ona sadrži dobro i da treba da se nadje u nekoj drugoj varijaciji.

Dvadeset prvi vek, poznatiji pod imenom ‘informaciono doba’, u mnogome je izmenio način pristupa različitim naučnim istraživanjima. Problemi, čija su rešenja bila nezamisliva, danas se rešavaju u mili sekundama i za kratko vreme mogu se otkriti latentne veze među podacima.

ZAHVALNICA

Rad na temu primene istraživanja podataka za otkrivanje obrazaca u domenu kulinarskih recepata deo je projekta u okviru nastavnog programa predmeta Napredne informacione tehnologije, na departmanu za Industrijsko ineženjerstvo i menadžment, na modulu za Informaciono-upravljački i komunikacioni sistemi. Profesor na predmetu i mentor ovog rada je prof. Dubravko Ćulibrk, docent na Fakultetu tehničkih nauka.

LITERATURA

[1] Fadi et. al Badra, "TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking," in 9th European Conference on Case-Based Reasoning, Trier, Germany, 2008.

[2] K, J Hammond, "A model of case-based planning," in Proc. 5th National Conf. on Artificial Intelligence, vol. 1, August, 1986., pp. 267-277.

[3] S Russell and P Norving, Artificial Intelligence, A Modern Approach.: Prentice Hall, 1994.

[4] A Hashimoto et al., "Smart kitchen: A user centric cooking support system," in Information Processing and Management of Uncertainty in Knowledge-Based Systems, June, 2008, pp. 848-854.

[5] Doman Keisuke, Cheng Ying Kuai, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase, "Video CooKing: Towards the Synthesis of Multimedia Cooking Recipes," in Advances in Multimedia Modeling - 17th International Multimedia Modeling Conference, Taipei, Taiwan, January 5-7, 2011, pp. 135-145.

[6] M Ohira, T Ozono, and T Shintani, "Implementing a recipe search system “minerecipe” using similarity-assessment knowledge," in 62nd Bi-Annual Convention, vol. 3, IPS Japan, 2011, pp. 129-130.

[7] K Ishihara, "An evaluation on the recommendation method for personal taste recipe," Technical Report, IEICE 2008.

[8] M Mori, K Kurihara, K Tsukada, and I Siio, "A system to enrich food color," Technical Repor t2007-80 2008.

[9] Rakesh Argawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules in large databases," in Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, September, 1994, pp. 487-499.

[10] G Webb and S Zhang, "Removing trivial association in association rule discovery," in Proceedings of the 1st International NAISO Congress on Autonomous Intelligent Systems (ICAIS), Geelong, Australia, 2002.

[11] Ian Witten, Frank Eibe, and Hall Mark, Data Mining: Practical machine learning tools and techniques, 3rd Edition. San Francisco, California: Morgan Kaufmann, 2011.

[12] R Agrawal and R Srikant, "Fast algorithms for mining association rules," in Proceedings of the 20th International Conference on Very Large Databases (VLDB), Santiago, Chile, 1994, pp. 487-499.

[13]J Han and M Kamber, Data Mining Concepts and Techniques.: Morgan Kaufmann, 2012.

[14] C Gyorödi and R Gyorödi, "Mining Association Rules in Large Databases," in Proceedings of of Oradea EMES, Oradea, Romania, 2002, pp. 45-50.

ABSTRACT

Abstract—**We are witnessing a boom in the number of online services that offer users a variety of recipes for the preparation of food and beverages. Systems that enable the automatic analysis of recipes, as well as the detection of patterns and principles that are specific to this data set reported a significant potential for applications and the opportunity to improve these services. This paper describes the application of algorithms for the detection of association rules to discover patterns which frequently occur in recipes. The study presented is based on publicly available recipes for the preparation of beer. Using data mining techniques we identified the most common ways in which the ingredients are used. The detected patterns can used to generate a new "universal" recipe for beer, which represents an optimized mixture of all kinds of beers and their methods of preparation.**

Key words : Data Mining; Association Rules; Apriori Algorithm

Data Mining and Pattern Discovery for Culinary Recipes

Nevena Nikolić, Kristina Pejić, Bernadeta Ralbovski