Специализированные методы визуализации корпусных данных

Бобров Евгений

Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Выпускная квалификационная работа бакалавра

Научный руководитель — к.ф.-м.н., Майсурадзе А.И.

7 июня 2017

Процесс анализа данных

В индустрии анализа данных все задачи — соответственно, методы и технологии их решения — принято подразделять на три уровня:

1. Моделирование данных и их интеграция в информационную систему;
2. Разработка аналитических пространств в целях оперативного анализа (OLAP, классическая визуализация);
3. Интеллектуальное (математическое) моделирование информационных процессов (машинное обучение, специализированная визуализация).

Полный цикл анализа данных должен решать задачи всех трех уровней.

Цели и задачи работы

Цели

- Проведение полного цикла анализа на всех трех уровнях;
- Преобразование исходных данных в удобное и понятное для человека представление.

Предметная область — научные исследования (наукометрия)
Исходные данные — корпус научных публикаций

Задачи

1. Реляционное моделирование слабоструктурированных данных из WoS;
2. Разработка и реализация интерактивной информационной панели;
3. Кластеризация и различная визуализация графов соавторства для стран и ключевых слов.

Сложная структура описания публикации

Computational Mathematics and Mathematical Physics

February 2017, Volume 57, Issue 2, pp 350–361

Aggregation of multiple metric descriptions from distances between unlabeled objects

Authors | Authors and affiliations

A. I. Maysuradze
M. A. Suvorov

1. Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia

Скалярные признаки

- Журнал
- Год

Множественные признаки

- Множество авторов
- Множество ключевых слов

Пары (научный центр, страна)

Коллекция научных статей

Authors	Title	Source	...
Marchand-J Pigeon-M Bager-D Talbot-C	Influence of Chloride Solution Concentration o...	ACI MATERIALS JOURNAL 1999, Vol 96, Iss 4, pp ...	...
Pauly-TM Lingvall-P	Effects of Mechanical Forage Treatment and Sur...	ACTA AGRICULTURAE SCANDINAVICA SECTION ...	...
...	...	...	...

- 364221 публикация
- 398972 авторов
- 100671 ключевых слов
- 69093 университетов
- 4386 журналов
- 195 стран
- 16 языков
- 14 типов документов

Концептуальная модель коллекции научных статей

Реляционное моделирование

Ненормализованное представление

О декомпозиции отношений по операции соединения

Теорема Хита

Если отношение $r(A, B, C)$ удовлетворяет функциональной зависимости $A \to B$, то оно равно соединению его проекций $r(A, B)$ и $r(A, C)$:
$(A \to B) \Rightarrow (r(A, B, C) = r(A, B)\ JOIN\ r(A, C))$

Пятая нормальная форма

Информационная панель

«Это экран, наполненный важной информацией, призванной достичь одну или несколько целей. Информация собрана и расположена на одном экране, поэтому сразу может быть воспринята человеком.» (Стефан Фью)

Многомерная модель строится автоматически по реляционной.

Граф соавторства

Содержательные задачи:

- Выделение научных сообществ — страна.
- Выделение научных тематик — ключевое слово.

Кратность ребра — количество совместных статей.

Мультиграф $G = (V, E)$. V — страны (слова). Страны v_i и v_j связаны ребром e_{ij}, если есть публикация a, такая что v_i и v_j входят в описание a. Одна публикация a превращается в C_n^2 рёбер, где n — число стран (слов) в описании статьи a.

Стохастические блоковые модели

Стохастическая блоковая модель — один из простейших генеративных процессов. Каждая вершина в одной группе: B — число групп, b_i — группа вершины v_i. Матрица смежности групп из $\mathbb{R}^{BxB}$ содержит элементы e_{rs}, равные числу рёбер между группами r и s.
Идея. Вершины графа принадлежат одной группе, если схожи вероятности быть соединёнными с остальными вершинами графа.

Стохастические блоковые модели

Пусть есть разбиение $b = \{b_i\}$ графа по B группам, где $b_i \in [1, B]$. Ещё понадобятся скрытые параметры модели θ. Тогда определена генеративная модель, описывающая граф G:

$$p(G, b, \theta)$$

Правдоподобие описывается разбиением b и параметрами θ:

$$p(G|b, \theta)$$

Априорная вероятность разбиения и параметров выбирается нами:

$$p(b, \theta)$$

Обоснованность определяется вероятностью:

$$p(G) = \sum_{\theta, b} p(G|b, \theta) p(b, \theta)$$

Стохастические блоковые модели

Апостериорная вероятность по формуле Байеса:

$$p(b|G) = \frac{p(G,b)}{p(G)} = \frac{\sum\limits_{\theta} p(G,b,\theta)}{p(G)} = \frac{\sum\limits_{\theta} p(G|b,\theta)p(b,\theta)}{p(G)}$$

Во многих моделях, в том числе в описываемой ниже, параметры θ жёстко зафиксированы методом максимального правдоподобия как $p(G|b,\hat{\theta}) = \max\limits_{\theta} p(G|b,\theta)$. Это позволяет существенно упростить модель к виду:

$$p(b|G) = \frac{p(G|b,\hat{\theta})p(b,\hat{\theta})}{p(G)},$$

формула может быть переписана в терминах *энергии*:

$$p(b|G) = \frac{\exp(-Q)}{p(G)}$$

тогда

$$p(b|G) \to \max \Leftrightarrow Q \to \min$$

$$Q = -\ln p(G|\hat{\theta},b) - \ln p(\hat{\theta},b) \to \min_{b}$$

Вложенная многоуровневая модель

Получится иерархическая кластеризация. Количество уровней и количество групп на каждом уровне определяется автоматически.

thailand
malaysia
u-arab-emirates
jordan
mozambique
oman
trinid-&-tobago
syria
yemen
liechtenstein
angola
bermuda
micronesia
turkey
chile
fr-polynesia
haiti
ukraine
moldova
kenya
uganda
gambia
senegal
malawi
mali
zambia
cameroon
papua-n-guinea
zaire
madagascar
congo
benin
cote-ivoire
niger
cambodia
bahrain
cent-afr-republ
myanmar
new-caledonia
rwanda
togo
bhutan
lesotho
solomon-islands
sierra-leone
somalia
liberia
bahamas
maldives
tonga
philippines
peru
indonesia
nigeria
burkina-faso
nepal
guatemala
paraguay
dominican-rep
jamaica
guinea
seychelles
brunei
grenada
swaziland
burundi
singapore
hong-kong
panama
tanzania
ghana
greenland
zimbabwe
guinea-bissau
sudan
mexico
slovenia
andorra
israel
serbia-monteneg
slovakia
ireland
vatican
hungary
india
south-korea
brazil
peoples-r-china
greece
czech-republic
england
germany
usa
portugal
australia
scotland
neth-antilles
austria
japan
norway
sweden
netherlands
italy
france
finland
denmark
poland
russia
belgium
canada
switzerland
spain
boston-univ.
romania
bulgaria
kazakhstan
uzbekistan
mongol-peo-rep
montenegro
cyprus
azerbaijan
armenia
ecuador
colombia
honduras
nicaragua
uruguay
venezuela
belize
eritrea
bolivia
costa-rica
iran
kyrgyzstan
guadeloupe
libya
namibia
botswana
cuba
morocco
pakistan
macedonia
luxembourg
laos
vanuatu
sri-lanka
ethiopia
vietnam
bangladesh
mauritania
st-lucia
malta
barbados

moldova
kenya
uganda
gambia
senegal
malawi
mali
zambia
cameroon
papua-n-guinea
zaire
madagascar
congo
benin
cote-ivoire
niger
cambodia
bahrain
cent-afr-republ
myanmar
new-caledonia
rwanda
togo
bhutan
lesotho
solomon-islands
sierra-leone
somalia
liberia
bahamas
maldives
tonga

usa
portugal
australia
scotland
neth-antilles
austria
japan
norway
sweden
netherlands
italy
france
finland
denmark
poland
russia
belgium
canada
switzerland
spain
boston-univ,
romania
bulgaria
kazakhstan
uzbekistan

Рис.: Карта мира, отражающая научное взаимодействие стран

evolution
time
disease
syndrome
resistance
serum
insulin
weight
glucose
infection
model
systems
dynamics
behavior
theory
temperature
size
stability
properties
models
modeling
design
flow
metabolism
amino
acids
pressure
plasma
food
responses
synthesis
degradation

system
binding
protein
diffusion
oxygen
absorption
chemistry
emission
light
films
transition
exchange
electron
ray
reaction
state
complexes
chemical
transfer
energy
metal
magnetic
crystal
spectroscopy
structure
liver
vitro
inhibitor

Заключение

На защиту выносится:

- Построение реляционной модели для коллекции научных статей;
- Разработка информационной панели для экспериментальных данных;
- Решение аналитических задачи по данным научных статей;
- Кластеризация и визуализация графов соавторства стран и ключевых слов в коллекции научных статей. Раскраска карты мира.

Кумулятивное распределение статей по годам
420000
400000
380000
360000
340000
320000
300000
280000
260000
240000
220000
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
article_year

Распределение статей по странам
country_name
sweden
denmark
finland
usa
germany
england
france
italy
netherlands
norway
spain
canada
russia
switzerland
japan
australia
belgium
Другие
0
20000
40000
60000
80000
100000
120000
140000
160000
180000
200000
220000

Распределение статей в наиболее популярных университетах
36000
34000
32000
30000
28000
26000
24000
22000
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
univ-helsinki
karolinska-inst
lund-univ
univ-copenhagen
uppsala-univ
univ-gothenburg
tech-univ-denmark
royal-inst-technol
univ-turku
umea-univ
university_name

Распределение статей по наиболее популярным авторам
johansson-b
andersson-m
larsson-a
johansson-m
tuomilehto-j
karlsson-m
holst-jj
olsen-j
andersson-s
peltonen-l
nielsen-j
hemminki-k
jonsson-p
nilsson-j
nielsen-s
nilsson-m
gustafsson-ja
johansson-a
sorensen-ht
kaprio-j
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50

Распределение числа ссылок в библиографии

cent-afr-republ
angola
solomon-islands
mauritania
iraq
grenada
vanuatu
fiji
niger
azerbaijan
laos
new-caledonia
guatemala
afghanistan
gabon
eritrea
cameroon
papua-n-guinea
malawi
nepal
belize
myanmar
rwanda
yemen
namibia
botswana
costa-rica
hong-kong
malagasy-republ
u-arab-emirates
yugoslavia
cambodia
mali
mongol-people-rep
tunisia
senegal
ghana
cote-ivoire
madagascar
bangladesh
egypt
macedonia
north-ireland
syria
monaco
brunei
uganda
latvia
armenia
serbia
lebanon
kenya
thailand
taiwan
venezuela
bosnia-&-herceg
benin
slovakia
romania
serbia-monteneg
vietnam
bolivia
malta
south-korea
libya
guadeloupe
india
byelorus
morocco
tanzania
chile
sri-lanka
ukraine
luxembourg
jamaica
zambia
algeria
indonesia
vatican
sudan
croatia
argentina
mauritius
nigeria
israel
pakistan
mexico
saudi-arabia
singapore
turkey
estonia
colombia
bahrain
lithuania
philippines
gambia
swaziland
burkina-faso
zimbabwe
ethiopia
iran
ecuador
french-guiana
tunisia
uzbekistan
rep-of-georgia
neth-antilles
trinid-&-tobago
uruguay
montenegro
peru
kuwait
mozambique
lesotho
panama
moldova
cuba
qatar
kazakhstan
guinea-bissau
nicaragua
albania
honduras
zaire
liechtenstein
jordan
paraguay
kyrgyzstan
syria
congo
dominican-rep
burundi
fr-polynesia
oman
guinea
w-ind-assoc-st
turkmenistan
tajikistan
bermuda
cook-islands
liberia

diffusion
design
properties
temperature
functional
multiple
oxygen
proteins
transition
messenger
magnetic
synthesis
membrane
inhibitor
behavior
electron
type
transport
field
mechanism
metal
films
mass
activity
chain
state
inhibition
electrophoresis
oxide
theory
complexes
apoptosis
activation
resonance
crystal
states
function
gamma
kinase
cross
high
energy
liver
family
vitro
stress
surface
reaction
ray
alpha
structure
dynamics
binding
density
brain
spectroscopy
modeling
mice
muscle
rat
phase
molecular
stability
chemical
infection
model
cells
receptor
flow
identification
transfer
chemistry
receptors
expression
protein
rna
analysis
systems
glucose
serum
factor
blood
system
exchange
randomized
women
water
single
cell
human
gene
growth
acid
escherichia
calcium
monoclonal
tumor
organic
trial
risk
induced
size
light
beta
disease
evolution
clinical
peptide
dna
cancer
and
food
carbon
coronary
hormone
lung
coli
mechanisms
emission
resistance
metabolism
syndrome
heart
antibodies
pressure
acids
absorption
nitrogen
forest
insulin
polymorphism
therapy
amino
dioxide
chromatography
tissue
models
factors
carcinoma
sequence
complex
children
mortality
breast
acute
plasma
population
responses
mutations
degradation
weight
time
epidemiology
survival
matter
age
sweden
bone
quality
exposure
life
health