Context Aware Predictive Analytics:  
Motivation, Potential, Challenges

Mykola Pechenizkiy

31 October 2011
University of Bournemouth, England
http://www.win.tue.nl/~mpechen/projects/CAPA
Outline

• What is context? In supervised learning tasks?
• Examples of why context is important in different applications
• Context-Aware Predictive Analytics framework
  – Motivation for multilevel model design & (re)training
  – Research questions and current directions
• CAPA in Web Analytics:
  – Advertisement, recommender systems, and alike;
  – Two major tasks: demand prediction (popularity) & relevance prediction (personal recommendations);
  – Importance of different data sources integration
What is Context?

• Definition of context depends on the context ...
  – Many (formal & vague) definitions in different fields
• Context as any *additional* information that
  – enhances the understanding of the instance of interest,
  – and in particular helps us to classify this instance
• We do not need _the_ formal definition of context in this talk;
  – instead we’ll consider several intuitive example
If we have seasonality, simple thresholding does not work; need to adjust w.r.t. the season.
Context in Change Detection

Online Mass Flow Prediction in CFB boiler:

Subtasks: identify change points and distinguish them from outliers

Both are difficult due to:
- asymmetric nature of the outliers
- short consumption periods within feeding stages

Contextual features characterize/categorize behavior of a product

\[ t = i - n \]
\[ t = i - 1 \]
\[ t = \text{now} \]

<table>
<thead>
<tr>
<th>( t )</th>
<th>History</th>
<th>Temp</th>
<th>Holiday</th>
<th>Promo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>{y(i-n) .. y(i-1)}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predict label of each new instance

http://dx.doi.org/10.1016/j.eswa.2011.07.078
Context in Classification

Not predictive alone but a subset of features with the contextual attribute(s) becomes (much) more predictive

Time of the day context

no context

https://sites.google.com/site/zliobaite/Zliobaite_PAKDD11_CR.pdf
Context-aware Driving Route Recognition

Baseline approach: LCSS matching on the database of existing reference routes

Time of day context
Context in Face Recognition

Variety of poses, lighting, facial expression, partial occlusions

- Finding a model or a single way to learn a model that recognizes face/identifies a person under such diverse settings is difficult
  - Context may not effect/identify an object, but can help to choose a better preprocessing technique, classifier etc

- Just adding background features to the representation space is not enough, i.e. multi-level prediction is needed

But e.g. “has glasses” can be a contextual attribute that e.g. suggests a model not to rely on the position of eyes, but can be also a predictive attribute/prior of belonging to a particular class

31 Oct 2011, U. Bournemouth

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Context in Word Sense Disambiguation

I went fishing for some sea bass. The bass line of the song is too weak.

• WSD - the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings (polysemy)
• (Un/Semi-)Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words
  – Which are predictive features?
  – Which are contextual?
Context in IR & RecSys

• User Context
  – Preferences, usage history, profiles

• Document/Product Context
  – Meta-data, content features

• Task Context
  – Current activity, location etc.

• Social Context
  – Leveraging the social graph
This can be refined further:

PHYSICAL CONTEXT
- location, time, associates, ...

USER’S OVERALL CURRENT TASK

INFORMATION NEED

USER’S KNOWLEDGE

USER’S OTHER TASKS

USER’S MODEL OF INFORMATION SEEKING

RELEVANCE FEEDBACK

USER’S SEARCH QUERY

INDEX FILE

RETRIEVED DOCUMENTS

Jones & Brown, 2004 The Role of Context in Information Retrieval
Mobile IR

• Profile, Location, Time,
• Activity, Agenda, Service,
• Preferences, Situation,
• Environment, Social Context

Context is certainly important …

… but it is not obvious what elements of context are useful, when and why!
Learning vs. Engineering CAPA

• Examples of engineering of context-aware/multi-level classification systems
  – Multi-lingual sentiment classification, emotion recognition
  – Relevant content identification in a web-page

• What can/should we design and what should we aim to induce from the data?
Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Multi-level decision making: seemingly more complicated process in fact simplifies the learning task by explicitly pointing out to the algorithm what to do.

Similar ideas are practiced a lot in many domains where we want to introduce the knowledge of the domain into the mining process, data fusion, information fusion.
Identifying Content Worth Indexing

1st: Find web page segments
2nd: Decide whether a segment contains relevant info

In a recent interview with Priyank Garg, Director of product management for Yahoo! Search Technology, conducted by Eric Enge, we were told that Yahoo breaks pages down into template sections to distinguish between noisy, or boilerplate content, and unique content:

One of the things Yahoo! has done is look for template structures inside sites so that we can recognize the boiler plate pages and understand what they are doing. And as you can expect, a boiler plate page like a contact us or an about us is not going to be getting a lot of anchor text from the Web and outside of your site. So there is natural targeting of links to your useful content.

We are also performing detection of templates within your site and the feeling is that that information can help us better recognize valuable links to users. We do that algorithmically, but one of the things we did last year around this time is we launched the robots-NoContent tag, which is a tool that webmasters can use to identify parts of their site that are actually not unique content for that page or that may not be relevant for the indexing of the page.

If you have ads on a page, or if you have navigation that’s common to the whole site, you could take more control over our efforts to recognize templates by marking those sections with the robots-NoContent tag. That will be a clear indicator to us that as the webmaster who knows this content, you are telling us this part of the page is not the unique main content of this page and don’t recall this page for those terms.
Supervised Learning

1. Training:
   \[ y = L(X) \]

2. Application:
   \[ y' = L(X') \]

Population (source)

Historical data

Model \( L \)

Testing data

Labels

\( x \)

\( y \)

\( x' \)

\( y' \)

Label?
Sup. Learning with Context-Awareness

Training:

??

Application:

\[ y' = L_j (X') \]
\[ L_j \leq G(X',E) \]

31 Oct 2011,
U. Bournemouth

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Design: (Re-)Learning Classifiers & Context

- Context Awareness
- Predictive Analytics
- Motivation, Potential, Challenges

Mykola Pechenizkiy, Eindhoven University of Technology
Application: Casting Context-Aware Predictions

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Monitoring Changes in Data, Context & Predictions

instances $X_{j-6} X_{j-5} X_{j-4} X_{j-3} X_{j-2} X_{j-1} X_j$

environment

PREDICTOR $L_i$

predictions $y_{j-6} y_{j-5} y_{j-4} y_{j-3} y_{j-2} y_{j-1} y_j$

monitoring over time change detection
Three Generic Phases Together
Research Questions

• How to define the context (form and maintain contextual categories) in web analytics?

• How to connect context with the prediction process in predictive web analytics?

• How to integrate change detection mechanisms into the prediction process in web analytics?

• How to ensure integration and feedback mechanisms between change detection and context awareness mechanisms?

• What should a reference architecture allowing to plug in new context aware prediction techniques for a collection of web analytics tasks look like?
Our goal: Framework to integrate context-awareness and change detection in Predictive Web Analytics

Two major types of tasks we are interested in
- Demand prediction
- Personalized recommendations

Modeling user behavior based on click-stream data
- Traffic data (page views, visits)
- Navigation data (traversed paths)
- Origin data (referrers, SE, keywords)
- System data (browser, resolution, plugins)

But besides click-stream data
- Customer data (location, lang., history, interests)
- Content data (webpages)

Contextual data (customer related, content related or environment related; calendar, weather, media highlights, twitter stream)
Context in Predictive Web Analytics

- Banner popularity and relevance predictions
  - Clickstream data and banner click data
- Social media and Web Analytics
  - How to link these two sources
- Information portals
  - Bounce rate reduction with recommenders
Framework for Studying Context-aware Banner Selection:

*Topical Online Portal Advertising*

Agus Wiro Susilo

Master project
Feb – Aug 2011
TU Eindhoven, Advertisement, Klikniews
Online advertising settings (stakeholders)

- = user
- = publisher
- = banner
- = ad server
- = advertiser
Context-aware advertising example

Context: summer holiday

Two sources for learning context from two separate systems:
- Banner click data
- Clickstream data

More relevant with the context
What we **can** learn:
- What and when banner is clicked
- Who clicks the banner

What we **cannot** learn:
- What other banners are shown with the clicked banner
- Clicking user behavior pattern
Context from clickstream data

What we can learn:
- User browsing behavior (browsing path)
- User segmentations (by locations, browsers, etc)
- Effectiveness of a page (FAQ, documentation, etc)

What we cannot learn:
- Banner click pattern (there is no banner click information)

- Access time
- Page URL
- Banners shown
- IP Address
- User Id
- Screen resolution
- Browser type
- Visit path
- etc.
Richer data, from which we can learn:
- What banners combinations produce more clicks
- What banners combinations do not produce clicks
- Preferred banners for new user
- Preferred banners for repeating user
- etc.
Matching Ads to Content Context

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Matching Ads to Content Context

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology

31 Oct 2011, U. Bournemouth
Matching to “Wrong” Context

Qantas emergency landing

Fly To Australia Cheap!
Save up to 85% on Australia Flights With Our Discounted Airfares!
Travelation.com

31 Oct 2011, U. Bournemouth

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Matching to “Wrong” Context
Matching to “Wrong” Context

Wildfires Rage in Texas; Thousands Evacuated

Firefighters try to control at least 63 wildfires across the Lone Star state – including one that has destroyed nearly 600 homes in Central Texas – as thousands of evacuees take shelter.

Source: AP/The Daily Sentinel
Matching to “Wrong” Context

31 Oct 2011, U. Bournemouth

Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Social Media Measurement Framework:

Support for Web Analytics and Search Engine Marketing

Murat Ongan

Master project
Feb – Aug 2011
TU Eindhoven, Adversitement, evaluation: Vodafone data
Web Analytics

- Measure and report web usage
  - Metrics and dimensions
Social Media Measurement

Sentiment analysis for coca-cola

Sentiment by Percent
- Positive (77%)
- Negative (23%)

Sentiment by Count
- Positive (22777)
- Negative (6942)

Zoom: 1d 5d 1m 3m 6m 1y  Max

Positive 79
Negative 29
May 31, 2011
# Natural Data Alignment

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Web Analytics</th>
<th>Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic</td>
<td>Page view</td>
<td>SM post</td>
</tr>
<tr>
<td>Unique</td>
<td>Visitor</td>
<td>Author</td>
</tr>
<tr>
<td>Success</td>
<td>Conversion</td>
<td>Positive message</td>
</tr>
<tr>
<td>Rate</td>
<td>Conversion rate</td>
<td>Percent positive</td>
</tr>
<tr>
<td>Date/time</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Location</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Medium</td>
<td>Direct, organic, referral</td>
<td>Twitter, news, blog</td>
</tr>
<tr>
<td>Source</td>
<td>Website domain</td>
<td>Twitter user / domain</td>
</tr>
<tr>
<td>Feature</td>
<td>Search keyword</td>
<td>Post content</td>
</tr>
</tbody>
</table>
Architecture Overview

Data Retriever → Database → Data Selector → GUI

Data Pre-processor

Web Analytics Integration

GUI → Data Analyst
(Context-aware)

Retrieval & Recommendation of Higher Education Programmes:

I Know Where You Will Study Next Year

Thijs Putman

Master project
Oct 2009 – May 2010
TU Eindhoven, MastersPortal

Found 5 programmes, displaying 1 - 5

Business Information Technology (MSc)
University of Twente (UT) – Enschede, Netherlands

As a Business Information Technology professional, you know that introducing a new information system is more than a technical installation, and requires insight in business processes and culture.

Time to change!
Master in Entrepreneurship and Innovation: enrol now!

Business Information Systems (MSc)
Eindhoven University of Technology (TU/e) – Eindhoven, Netherlands

The Master's degree program in Business Information Systems combines computer science and business management. The program places the emphasis on the development of high-quality information systems based on a business perspective.

IMMIT: International Master in Management of Information Technology (MSc)
Tilburg University – Tilburg, Netherlands

Information and information technology (IT) are increasingly an integral part of products and services as well as the foundation of business processes. Organisations must know how to make the right choices with respect to new IT systems and at the...

IMSE: International Master in Service Engineering, Erasmus Mundus (MSc)
Tilburg University – Tilburg, Netherlands

Information and information technology (IT) are increasingly an integral part of products and services as well as the foundation of business processes. Organisations must know how to make the right choices with respect to new IT systems and at the...
Impact of Google on MastersPortal.eu

- Google’s share steady around 50% since March 2009
Impact of Google on MastersPortal.eu

• Visitors are often referred to pages with detailed info

Google

Business Information Systems, MSc
Eindhoven University of Technology (TU/e), Mathematics and Computer Science

Disciplines:
- Computer Science & IT
- Engineering & Business
- Informatics & Information Science

Quick facts

<table>
<thead>
<tr>
<th>Country</th>
<th>Netherlands</th>
<th>Duration</th>
<th>24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Eindhoven</td>
<td>Start Date</td>
<td>September</td>
</tr>
<tr>
<td>Educational Form</td>
<td>Taught</td>
<td>Languages</td>
<td>English</td>
</tr>
<tr>
<td>Education Variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Tuition Fee</td>
<td>€ 1820</td>
<td></td>
<td>€ 6600 (non-EEA)</td>
</tr>
</tbody>
</table>

Programme Description

Today’s business world is unthinkable without the major contribution made by computer science. Information and communication technology (ICT), and especially information systems, have become a cornerstone of business management.

• Many visitors leave after first page
  – Leads to a high bounce-rate
    • Bounce-rate for Google 84%; Other, non search-engine, referrals: 65%
Solution: Programme Recommender

• Provide a set of "Related Master’s Programmes"

• Based on the academic contents of each Master’s programme
  – Select from the over 14,000 programmes in the database
  – Mostly unstructured information
Business Information Systems, MSc

Eindhoven University of Technology (TU/e), Mathematics and Computer Science

**Disciplines:**
- Computer Science & IT
- Engineering & Business
- Informatics & Information Science

**Quick facts**

- **Country:** Netherlands
- **City:** Eindhoven
- **Duration:** 24 Months
- **Starting Date:** September
- **Education Variants:**
  - Part Time
  - Full Time
- **Educational Form:** Taught
- **Languages:** English
- **Annual Tuition Fee:**
  - €1620
  - €8600 (non-EEA)

**Programme Description**

Today’s business world is unthinkable without the major contribution made by computer science. Information and communication technology (ICT), and especially information systems, have become a cornerstone of business management in multinationals, in banks and insurance companies, and in small and medium-size enterprises. Companies have become dependent on these increasingly complex systems, and out of necessity place stringent demands on their reliability and security.

The Master’s degree program in Business Information Systems combines computer science and business management. The program places the emphasis on the development of high-quality information systems based on a business perspective.

As a graduate of this program you will combine a scientific attitude with a model-driven engineering approach. You will be able to understand the demands that are placed on information systems, and to initiate and implement new applications. This approach can already be seen in the compulsory courses of the program. The compulsory computer science courses are Software architecture, Web information systems, Database models, Process modeling and Information retrieval. The compulsory business management courses are Information management, IT governance, E-business architecture and systems, Workflow management systems and Supply chain logistics and information management.

During the program you can place the emphasis on the computer science aspects or the business management aspects.
Deploy & Evaluate Two Alternatives

- *state-of-the-art recommendation approaches*
  1. Content-Based
Deploy & evaluate two alternatives

- *state-of-the-art* recommendation approaches
  1. Content-Based
  2. Collaborative
Experiment Results of Live Testing

Bounce-rate *

<table>
<thead>
<tr>
<th>Recommender Implementation</th>
<th>Bounce-rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>90.29%</td>
</tr>
<tr>
<td>Random</td>
<td>90.08%</td>
</tr>
<tr>
<td>Baseline</td>
<td>85.43%</td>
</tr>
<tr>
<td>Collaborative</td>
<td>82.03%</td>
</tr>
<tr>
<td>Content</td>
<td>81.49%</td>
</tr>
</tbody>
</table>

- Both approaches perform equally well
- Almost **doubles** visitor retention
Contextual Factors

• Categorising the visitor
  – Create groups of similar visitors; **visitor segments**

• Is performance **consistent** across different **visitor segments**?
An Adaptive Recommender System

- Content-Based Recommender
- Collaborative Recommender
A/B Live Testing

• An important element of R&D infrastructure
  – MastersPortal
  – Klicknews
  – Adversitement
  Web Analytics
  solution
• Two versions of the system
  - Splitting the traffic between them
  - Compute which one is doing better
Planned and Ongoing Work

Together with Adversitement & Klikniews

• Finding (anti)catalyst banners
  • Sports fits politics, but not fashion
• Study A/B testing infrastructure
  • experiment design, mining the results from the controlled experiments
• Matching Ads to Content Context
  • Also wrt positive–negative content
• Analysing different problems formulations: e.g. contract items as constraints (number of impressions to be guaranteed)
Context Aware Predictive Analytics: Motivation, Potential, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology