Implementation of a Portal Dedicated to Higgs Bosons for Experts and the General Public

Czech Technical University in Prague, Institute of Experimental and Applied Physics

André Sopczak on behalf of the International Particle Physics Outreach Group (IPPOG)

Introduction

- The implementation of a web portal dedicated to Higgs boson research is presented.
- A database is created with more than 1000 relevant articles using CERN Document Server API and web scraping methods.
- The database is automatically updated when new results on the Higgs boson become available.
- Using natural language processing, the articles are categorised according to properties of the Higgs boson and other criteria.
- The process of designing and implementing the Higgs Boson Portal (HBP) is described in detail.
- The components of the HBP are deployed to CERN Web Services using the OpenShift cloud platform.
- The web portal is operational and freely accessible on http://cern.ch/higgs.

Higgs Boson Research

- 1989 – 2000: CERN – Large Electron-Positron Collider
  - ALEPH
  - DELPHI
  - L3
  - OPAL
- 1987 – 2011: Fermilab – Tevatron Collider
  - CDF
  - D0
- 2010 – present: CERN – Large Hadron Collider
  - ATLAS
  - CMS

Research Resources

- 1000+ scientific publications (experimental results)
- Various types of experiments
- Various publishing methods
- New results each week/day
- Large number of articles – important to create a categorisation system

Goals

- Easy access to publications
- Collection and categorization
- Visualisation of development precisions
- Bringing the research closer to the public

Collecting data

Which data?

- Publications – title, abstract, tables, graphs.
- Measured values – masses, production modes, decay modes. . .

How?

- Fermilab (old websites) – web scraping
- CERN – CERN Document Server API
- Measured values – extract from text

Categorisation

- Goal of the publication
  - Experimental measurement
  - Search for “new physics”
- Observed events
  - Higgs boson production
  - Higgs boson decay
- Other properties
  - Number of collisions (luminosity)
  - Collision energy
  - Experiment
  - Current stage (preliminary, submitted, published)

Natural Language Processing

Vocabulary varies by the type of experiment

- Standard Model vs “new physics”
- Naive Bayes classifier
- Relies on the frequency of certain words
- No need to specify the words beforehand
- Training examples
- Tokenization, stopwords, lemmatization
- Python, scikit-learn, nltk

Numeric values and keywords detection

- Plain searching – ineffective
- Named Entity Recognition
- English corpus
- AI learns to recognize written text patterns
- Further algorithmic processing

NLP example of article title:

Search for charged Higgs bosons produced via vector boson fusion and decaying into a pair of W and Z bosons using proton-proton collisions at $\sqrt{s}$ = 13TeV.

Application

- Categorisation model
- Manual training
- Training and testing set ∼100 articles
- Standard F1-score

Categorization results

- True Positives: TP
- False Positives: FP
- False Negatives: FN

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

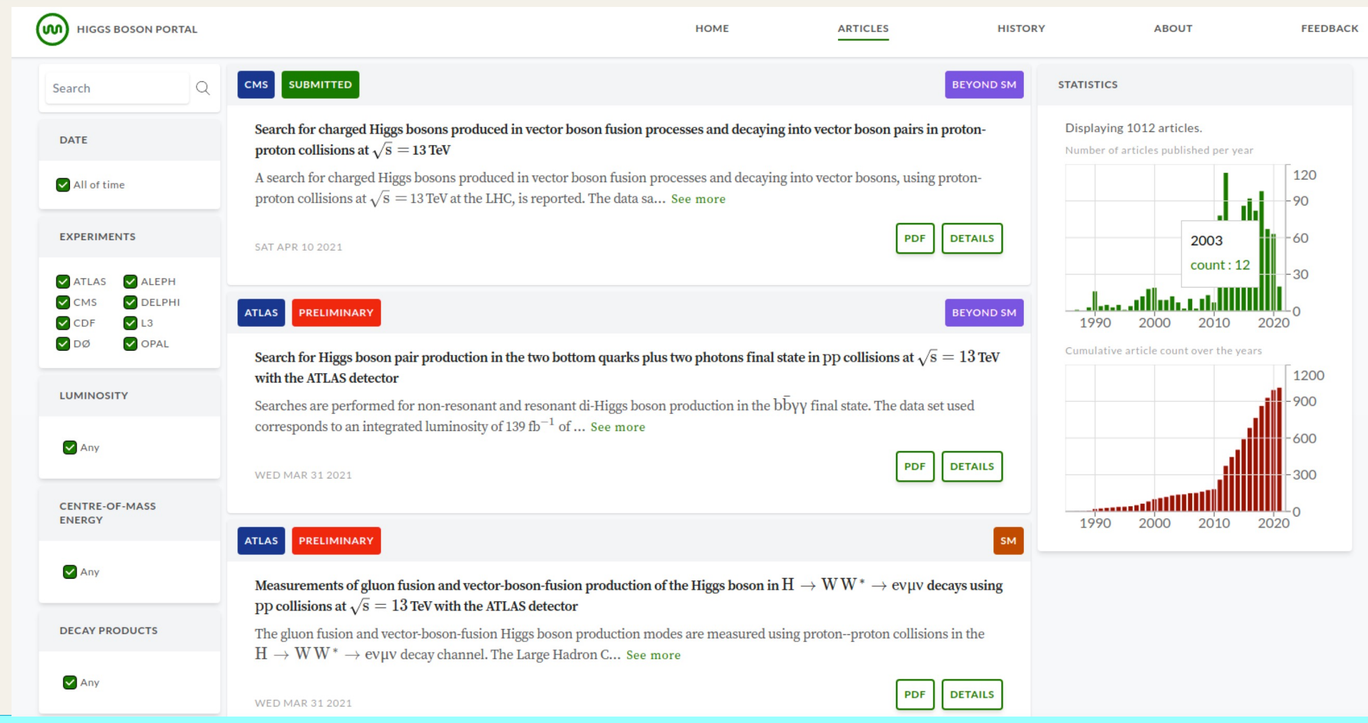
F1 score = 2 P*R/(P+R)

Category	Precision (%)	Recall (%)	F_1-score (%)
Luminosity	96	88	92
Energy	100	85	92
Production mode	87	85	86
Decay mode	81	79	80

Web Application

- User Interface and administration
- Administrator can adjust NLP categorization
- Categorised publications stored in a database – MongoDB
- Daily updates – Python cron jobs
- API – Flask
- UI and administration – React.js, Tailwind.css

User Interface



References/Acknowledgements

- Martin Kupka, CERN-THESIS-2020-053, Feasibility Study of a Portal to Provide Knowledge about Higgs Bosons to the General Public and Experts
- Peter Zacik, CERN-THESIS-2021-080, Implementation of a Portal Dedicated to Higgs Bosons for Experts and the General Public
- Antoine Vauterin, André Sopczak, 22nd IPPOG meeting, 17-19 Nov. 2021, https://indico.cern.ch/event/1084892, New Web-based Educational Tool for ATLAS
- 18th International Masterclasses 2022, https://physicsmasterclasses.org

The project is supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project number LTT 17018.