Automatic Metadata Extraction
The High Energy Physics Use Case

Joseph Boyd

École Polytechnique Fédérale de Lausanne

joseph.boyd@epfl.ch

August 20, 2015

Motivation

Aims

Outline

1 Introduction

2 Theory

3 Automatic Metadata Extraction

4 Data, Methods, and Implementation

5 Key Results

6 Conclusions

Why CRFs?

Mathematical Formulation

Solution Approach

Outline

1. Introduction
2. Theory
3. Automatic Metadata Extraction
4. Data, Methods, and Implementation
5. Key Results
6. Conclusions

Metadata Extraction

GROBID

GROBID - CRF Cascade

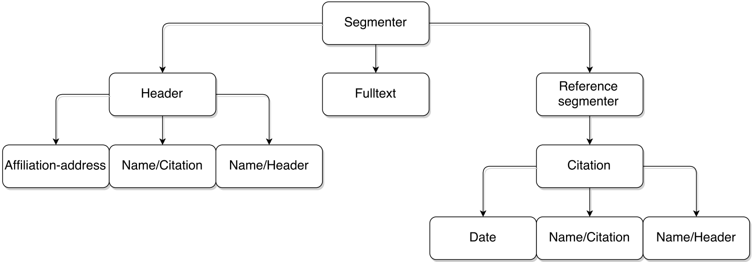
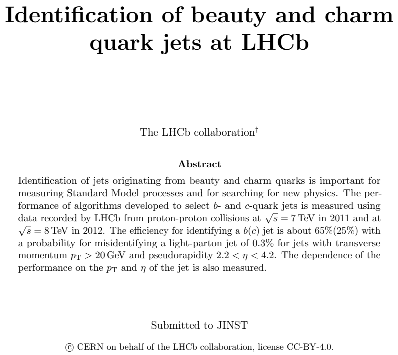


Figure: Cascade of models used by Grobid

Outline

1. Introduction
2. Theory
3. Automatic Metadata Extraction
4. Data, Methods, and Implementation
5. Key Results
6. Conclusions



Identification of beauty and charm quark jets at LHCb

The LHCb collaboration†

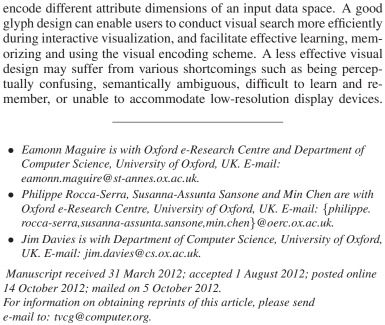
Abstract

Identification of jets originating from beauty and charm quarks is important for measuring Standard Model processes and for searching for new physics. The performance of algorithms developed to select b- and c-quark jets is measured using data recorded by LHCb from proton-proton collisions at $\sqrt{s} = 7\,\text{TeV}$ in 2011 and at $\sqrt{s} = 8\,\text{TeV}$ in 2012. The efficiency for identifying a $b(c)$ jet is about 65%(25%) with a probability for misidentifying a light-parton jet of 0.3% for jets with transverse momentum $p_T > 20\,\text{GeV}$ and pseudorapidity $2.2 < \eta < 4.2$. The dependence of the performance on the p_T and η of the jet is also measured.

Submitted to JINST

© CERN on behalf of the LHCb collaboration, license CC-BY-4.0.

(a) Collaboration field in header section.

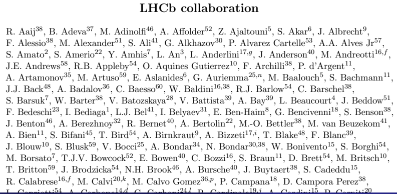


encode different attribute dimensions of an input data space. A good glyph design can enable users to conduct visual search more efficiently during interactive visualization, and facilitate effective learning, memorizing and using the visual encoding scheme. A less effective visual design may suffer from various shortcomings such as being perceptually confusing, semantically ambiguous, difficult to learn and remember, or unable to accommodate low-resolution display devices.

- *Eamonn Maguire is with Oxford e-Research Centre and Department of Computer Science, University of Oxford, UK. E-mail: eamonn.maguire@st-annes.ox.ac.uk.*
- *Philippe Rocca-Serra, Susanna-Assunta Sansone and Min Chen are with Oxford e-Research Centre, University of Oxford, UK. E-mail: {philippe.rocca-serra,susanna-assunta.sansone,min.chen}@oerc.ox.ac.uk.*
- *Jim Davies is with Department of Computer Science, University of Oxford, UK. E-mail: jim.davies@cs.ox.ac.uk.*

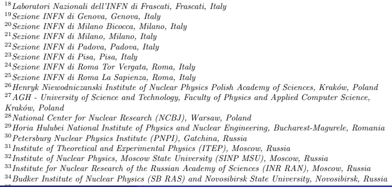
Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

(b) Discontinuous header data.



LHCb collaboration

R. Aaij[38], B. Adeva[37], M. Adinolfi[46], A. Affolder[52], Z. Ajaltouni[5], S. Akar[6], J. Albrecht[9], F. Alessio[38], M. Alexander[51], S. Ali[41], G. Alkhazov[30], P. Alvarez Cartelle[53], A.A. Alves Jr[57], S. Amato[2], S. Amerio[22], Y. Amhis[7], L. An[3], L. Anderlini[17,g], J. Anderson[40], M. Andreotti[16,f], J.E. Andrews[58], R.B. Appleby[54], O. Aquines Gutierrez[10], F. Archilli[38], P. d'Argent[11], A. Artamonov[35], M. Artuso[59], E. Aslanides[6], G. Auriemma[25,n], M. Baalouch[5], S. Bachmann[11], J.J. Back[48], A. Badalov[36], C. Baesso[60], W. Baldini[16,38], R.J. Barlow[54], C. Barschel[38], S. Barsuk[7], W. Barter[38], V. Batozskaya[28], V. Battista[39], A. Bay[39], L. Beaucourt[4], J. Beddow[51], F. Bedeschi[23], I. Bediaga[1], L.J. Bel[41], I. Belyaev[31], E. Ben-Haim[8], G. Bencivenni[18], S. Benson[38], J. Benton[46], A. Berezhnoy[32], R. Bernet[40], A. Bertolin[22], M.-O. Bettler[38], M. van Beuzekom[41], A. Bien[11], S. Bifani[45], T. Bird[54], A. Birnkraut[9], A. Bizzeti[17,i], T. Blake[48], F. Blanc[39], J. Blouw[10], S. Blusk[59], V. Bocci[25], A. Bondar[34], N. Bondar[30,38], W. Bonivento[15], S. Borghi[54], M. Borsato[7], T.J.V. Bowcock[52], E. Bowen[40], C. Bozzi[16], S. Braun[11], D. Brett[54], M. Britsch[10], T. Britton[59], J. Brodzicka[54], N.H. Brook[46], A. Bursche[40], J. Buytaert[38], S. Cadeddu[15], R. Calabrese[16,f], M. Calvi[20,k], M. Calvo Gomez[36,p], P. Campana[18], D. Campora Perez[38],



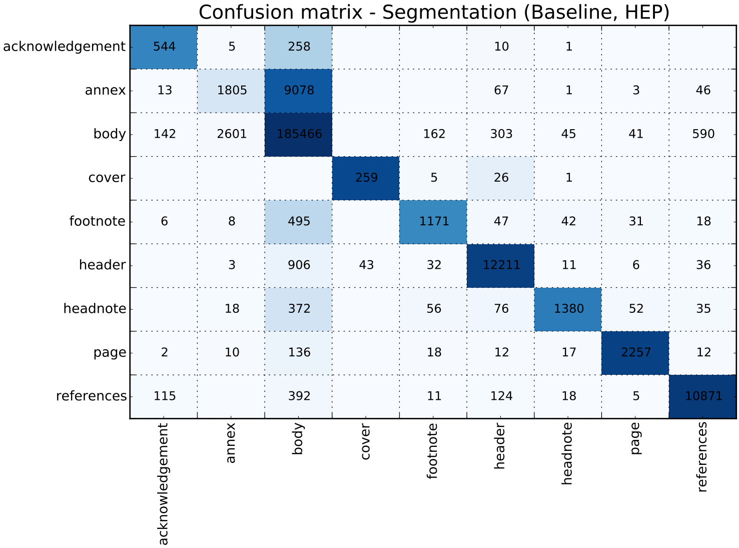
[18] *Laboratori Nazionali dell'INFN di Frascati, Frascati, Italy*
[19] *Sezione INFN di Genova, Genova, Italy*
[20] *Sezione INFN di Milano Bicocca, Milano, Italy*
[21] *Sezione INFN di Milano, Milano, Italy*
[22] *Sezione INFN di Padova, Padova, Italy*
[23] *Sezione INFN di Pisa, Pisa, Italy*
[24] *Sezione INFN di Roma Tor Vergata, Roma, Italy*
[25] *Sezione INFN di Roma La Sapienza, Roma, Italy*
[26] *Henryk Niewodniczanski Institute of Nuclear Physics Polish Academy of Sciences, Kraków, Poland*
[27] *AGH - University of Science and Technology, Faculty of Physics and Applied Computer Science, Kraków, Poland*
[28] *National Center for Nuclear Research (NCBJ), Warsaw, Poland*
[29] *Horia Hulubei National Institute of Physics and Nuclear Engineering, Bucharest-Magurele, Romania*
[30] *Petersburg Nuclear Physics Institute (PNPI), Gatchina, Russia*
[31] *Institute of Theoretical and Experimental Physics (ITEP), Moscow, Russia*
[32] *Institute of Nuclear Physics, Moscow State University (SINP MSU), Moscow, Russia*
[33] *Institute for Nuclear Research of the Russian Academy of Sciences (INR RAN), Moscow, Russia*
[34] *Budker Institute of Nuclear Physics (SB RAS) and Novosibirsk State University, Novosibirsk, Russia*

Model	HEP	CORA
Header	157 papers	**2506 papers**
Segmentation	**169 papers**	125 papers

Table: Number of training instances for each model from each dataset.

Outline

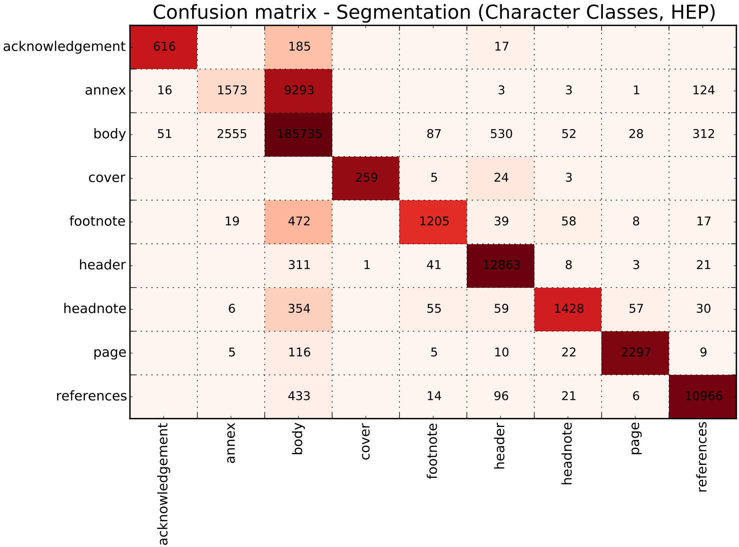
1. Introduction
2. Theory
3. Automatic Metadata Extraction
4. Data, Methods, and Implementation
5. Key Results
6. Conclusions



Confusion matrix - Segmentation (Baseline, HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	544	5	258			10	1		
annex	13	1805	9078			67	1	3	46
body	142	2601	185466		162	303	45	41	590
cover				259	5	26	1		
footnote	6	8	495		1171	47	42	31	18
header		3	906	43	32	12211	11	6	36
headnote		18	372		56	76	1380	52	35
page	2	10	136		18	12	17	2257	12
references	115		392		11	124	18	5	10871

Figure: Baseline confusion segmentation



Confusion matrix - Segmentation (Character Classes, HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	616		185			17			
annex	16	1573	9293			3	3	1	124
body	51	2555	185735		87	530	52	28	312
cover				259	5	24	3		
footnote		19	472		1205	39	58	8	17
header			311	1	41	12863	8	3	21
headnote		6	354		55	59	1428	57	30
page		5	116		5	10	22	2297	9
references			433		14	96	21	6	10966

Figure: Classes confusion segmentation

Outline

1 Introduction

2 Theory

3 Automatic Metadata Extraction

4 Data, Methods, and Implementation

5 Key Results

6 Conclusions