

josedevEZas

information retrieval and data science



contact

Porto, Portugal

joseleisdevEZas@gmail.com

<http://www.josedevEZas.com>

LinkedIn: jldevEZas

languages

Portuguese (native)

English (fluent)

programming

♥ Python, R, SQL,
Java, C#, Scala, Rust,
Gremlin, HTML, CSS,
JavaScript

education

2016–Now

Doctoral Program in Computer Science (MAP-i)

U.Porto, UA, UMinho

Classification (1st year): 19 (out of 20)

Graph-Based Entity-Oriented Search

Thesis statement:

A graph-based joint representation of unstructured and structured data has the potential to unlock novel ranking strategies, that are, in turn, able to support the generalization of entity-oriented search tasks and to improve overall retrieval effectiveness by incorporating explicit and implicit information derived from the relations between text found in corpora and entities found in knowledge bases.

During the course of this doctoral work, two main systems were developed: ANT, an entity-oriented search engine for the University of Porto (prototype); and Army ANT, a workbench for innovation in entity-oriented search.

<https://ant.fe.up.pt/>

<https://github.com/feup-infolab/ant>

<https://github.com/feup-infolab/army-ant>

<https://github.com/feup-infolab/army-ant-install>

<https://hub.docker.com/repository/docker/jldevEZas/army-ant>

2003–2010

Master in Informatics and Computing Engineering

U.Porto

Classification: 14 (out of 20)

Link Ecosystem of the Portuguese Blogosphere

This thesis explored the blogosphere as an ecosystem of blogs comprised not only of a set of posts and their individual content, but also of the interactions established through the hyperlinks connecting the posts. Based on this study, we discovered a correlation between the blog's centrality and the length and number of posts.

Thesis Classification: 18 (out of 20)

2000–2003

Technological Specialization in Informatics

Colégio de Gaia

Classification: 18 (out of 20)

experience

2016–2020

CSIG, INESC TEC

Porto, Portugal

Researcher / PhD Student

MAP-i PhD student and FCT grant holder, working on entity-oriented search. Former FourEyes researcher with a focus on studying information consumption on social media. Provided support to the research group through data engineering, by developing the infrastructure for Twitter data collection and analysis, as well as news collection and archiving.

2015–2020	INFORMATION SYSTEMS LABORATORY, UNIVERSITY OF PORTO <i>Researcher / PhD Student</i> MAP-i PhD student and FCT grant holder, working on entity-oriented search. Research and development of the ANT entity-oriented search engine. Particularly focused on query analysis methodologies, as well as the construction of specialized indexes and ontologies for efficient and effective question-answering through contextual widgets. The produced system was also aimed at providing a platform able to support the academic community with the experimentation of search engines backed by linked data.	Porto, Portugal
2017–2018	FACULTY OF ENGINEERING, UNIVERSITY OF PORTO <i>Invited Assistant</i> Lectured one semester on social network analysis, for the Information Management in Social Networks course, at the Master in Information Science, and one semester on relational and document databases, for the Databases course, at the Master in Informatics and Computing Engineering.	Porto, Portugal
2014–2015	INTERRELATE <i>Software Engineer / Data Scientist</i> <i>Data Science:</i> <ul style="list-style-type: none"> • Collection and analysis of several types of data, mainly originating from the web, particularly from social media. • Highly proficient in R, Python and SQL (from data collection to reporting), including topic tracking, outlier and anomaly detection, statistical smoothing, regression, classification, clustering and network analysis. • Skilled in dynamic reporting. • Developed several data visualization widgets. • Analyzed social networks and other graphs. <i>Software Engineering:</i> <ul style="list-style-type: none"> • Software design of web applications and analytics pipelines. • Technology evaluation and selection. • Supporting the team with new technologies or languages. • INTERRELATE Backoffice: <ul style="list-style-type: none"> – Devising strategies for automatic and periodical data collection and information extraction from unstructured user-generated content; – Systems integration, including team coordination (4 person project). <i>Programming:</i> <ul style="list-style-type: none"> • Worked in nearly every company project, using various different languages. • Developed several web applications and web services. • INTERRELATE Backoffice: <ul style="list-style-type: none"> – Web crawler development; – Development of data visualization widgets for internal statistics; – Front-end design and development. <i>Sysadmin:</i> <ul style="list-style-type: none"> • Manage the company servers, including security, backups and product deploys. • Maintain firewall rules and SMTP server. • DNS configurations (including SPF, for SPAM control). • Nginx configurations. • Company services maintenance (internal and external). • Log monitoring and bug reporting. • Real-time e-mail log error monitoring for critical systems and web service mobile monitoring for downtime or connection issues. • Support Ticket System configuration. • PayPal payment services. 	Porto, Portugal

2012–2013	SAPO LABS, UNIVERSITY OF PORTO <i>Researcher / Developer</i> Research and development of music discovery and recommendation techniques supported on a graph-based neighborhood approach, as well as on latent factor models based on matrix factorization algorithms. Development of methodologies to combine knowledge from different dimensions, including not only user profile, content and context features, but also social features, such as community structure. Related tasks required work on data mining, multimodal network analysis and community detection, and information retrieval.	Porto, Portugal
2011–2012	CRACS, INESC TEC <i>Researcher / Developer</i> Enhancement, adaptation and implementation of state of the art algorithms for network analysis, data mining and data visualization, in the context of intelligent information systems for cyber journalism, as part of the Breadcrumbs Project (UTA-Est/MAI/0007/2009).	Porto, Portugal
2010–2011	SAPO LABS, UNIVERSITY OF PORTO <i>Researcher / Developer</i> Work in the areas of data mining, network analysis, community detection, information retrieval, real-time data analytics and data visualization.	Porto, Portugal
2002–2003	NPF PORTUGUESE FREEBSD GROUP - npf.pt.freebsd.org <i>Documentation / Programming</i> Translation of the official FreeBSD documentation to portuguese, creation of how-to guides in portuguese, porting open source applications to FreeBSD and programming and content management of a hints widget.	Coimbra, Portugal

awards

2012	Certificate of Merit The 2012 IAENG International Conference on Data Mining and Applications Awarded to the best papers published in the conference.
------	--

communication skills

2020	Demo European Conference on Information Retrieval Prepared a video demonstration of the Army ANT system, which was exhibited during the online event, with a slot reserved for questions.
2020	Doctoral Consortium European Conference on Information Retrieval Participated in the doctoral consortium, presenting graph-based entity-oriented search as a unified framework in information retrieval.
2020	Lecture Software Systems Architecture course, at the Master in Informatics and Computing Engineering Described ANT search engine's system architecture, focusing on the requirements of an entity-oriented search engine.
2019	Presentation Symposium on Languages, Applications and Technologies Presented the graph-of-entity representation and retrieval model for entity-oriented search over combined data.
2019	Invited Speaker Creative CoLAB 2019 Introduced several concepts about hypergraphs, arguing about their generality, and presented an approach to mapping terms and entities with hypergraphs, which could be used to universally solve multiple entity-oriented search tasks.

- 2018 **Lecture** Information Description, Storage and Retrieval course, at the Master in Informatics and Computing Engineering
Described ANT search engine's system architecture, introducing several technical aspects of the building of an entity-oriented search engine.
- 2018 **Workshop** IEEE Student and Young Professional Congress
Presented ANT, focusing on its query understanding algorithm, based on the Score Hypergraph, as well as and Army ANT, showing how it can be used to research entity-oriented search.
- 2017 **Presentation** Symposium on Languages, Applications and Technologies
Presented a complete pipeline, from data acquisition, passing through information extraction and the automatic construction of knowledge base, to an information retrieval implementation.
- 2013 **Presentation** DEI, FEUP
Did a short presentation on the Juggle project for the department, in the context of DEI Talks.
- 2013 **Presentation** SAPO
Presented the Juggle project, including the developed graph-based recommendation algorithms, for the SAPO technical team.
- 2012 **Presentation** International Conference on Knowledge Discovery and Information Retrieval
Presented an interactive news clips visualization tool for applications in journalistic research and knowledge discovery.
- 2012 **Presentation** IAENG International Conference on Data Mining and Applications
Presented a methodology, based on community detection in multidimensional networks, for the creation of news context from a folksonomy of web clipping.
- 2011 **Poster** International AAAI Conference on Weblogs and Social Media
Presented research work on using the h-index to estimate blog authority.
- 2010 **Presentation** SAPO Labs, University of Porto
Presented Ciclope project for the Portugal Telecom's CEO and the media, as part of the inaugural presentation of Laboratório SAPO/U.Porto.
- 2010 **Poster** KDD Workshop on Social Media Analytics
Presented the blog popularity study conducted during my Masters degree.

interests

professional: search engines, data analysis, web development, blockchain, DeFi, algorithms and data structures, data visualization **personal:** investment, crypto, games, music, tv, movies

more information: <http://josedevezas.com/>

collaborations

- 2020 **Member of the program committee**
Graphs and More Complex Structures for Learning and Reasoning, Workshop At AAAI 2021
<https://sites.google.com/view/gclr2021/>
- 2019–2020 **Member of the program committee**
European Conference on Information Retrieval 2020 and 2021
<https://www.ecir2020.org/>, <https://www.ecir2021.eu/>
- 2017–2019 **Member of the program committee**
Complex Networks 2017, 2018 and 2019
<http://complexnetworks.org>

- 2019/2020 **Master's thesis co-supervision**
Building a domain-specific search engine that explores football-related search patterns
<https://hdl.handle.net/10216/128526>
- 2017/2018 **Participation in TREC 2018 with the University of Alicante**
FEUP at TREC 2018 Common Core Track – Reranking for Diversity using Hypergraph-of-Entity and Document Profiling
<https://trec.nist.gov/pubs/trec27/papers/FEUP-CC.pdf>
- 2016/2017 **Master's thesis co-supervision**
Named entity extraction from Portuguese web text
<http://hdl.handle.net/10216/106094>
- 2015/2016 **Master's thesis co-supervision**
Exploring the Sea: Heterogeneous Geo-Referenced Data Repository
<http://hdl.handle.net/10216/85612>

publications

journal articles

- Characterizing the hypergraph-of-entity and the structural impact of its extensions
 José Devezas and Sérgio Nunes
Applied Network Science - Special Issue of the 8th International Conference on Complex Networks and Their Applications 5.1 (2020) p. 79. 2020, DOI: 10.1007/s41109-020-00320-z
- Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge
 José Devezas and Sérgio Nunes
Open Computer Science 9.1 (June 2019) pp. 103–127. 2019, DOI: 10.1515/comp-2019-0006
- Graph-Based Entity-Oriented Search: Imitating the Human Process of Seeking and Cross Referencing Information
 José Devezas and Sérgio Nunes
ERCIM News. Special Issue: Digital Humanities 111 (Oct. 2017) pp. 13–14. 2017, URL: <https://ercim-news.ercim.eu/en111/special/graph-based-entity-oriented-search-imitating-the-human-process-of-seeking-and-cross-referencing-information>
- The community structure of a multidimensional network of news clips
 José Luís Devezas and Álvaro Reis Figueira
Int. J. Web Based Communities 9.3 (2013) pp. 411–429. 2013, DOI: 10.1504/IJWBC.2013.054911
- Finding Language-Independent Contextual Supernodes on Coreference Networks
 José Devezas and Álvaro Figueira
IAENG International Journal of Computer Science 39.2 (2012) pp. 200–207. 2012, URL: http://www.iaeng.org/IJCS/issues_v39/issue_2/IJCS_39_2_07.pdf

conference articles

- Graph-Based Entity-Oriented Search: A Unified Framework in Information Retrieval
 José Devezas
Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, 2020, DOI: 10.1007/978-3-030-45442-5_78
- Army ANT: A Workbench for Innovation in Entity-Oriented Search
 José Luís Devezas and Sérgio Nunes
Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, 2020, DOI: 10.1007/978-3-030-45442-5_56
- Characterizing the Hypergraph-of-Entity Representation Model

José Devezas and Sérgio Nunes

Complex Networks and Their Applications VIII - Volume 2 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019, 2019, DOI: 10.1007/978-3-030-36683-4_1

Graph-of-Entity: A Model for Combined Data Representation and Retrieval

José Devezas and Sérgio Nunes

Proceedings of the 8th Symposium on Languages, Applications and Technologies (SLATE 2019), 2019, Vila do Conde, Portugal, DOI: 10.4230/OASlcs.SLATE.2019.1

FEUP at TREC 2018 Common Core Track - Reranking for Diversity using Hypergraph-of-Entity and Document Profiling

José Luís Devezas, Sérgio Nunes, Antonio Guillén, Yoan Gutiérrez, and Rafael Muñoz

Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018, 2018, URL: <https://trec.nist.gov/pubs/trec27/papers/FEUP-CC.pdf>

Social Media and Information Consumption Diversity

José Devezas and Sérgio Nunes

Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018, 2018, URL: <http://ceur-ws.org/Vol-2079/paper5.pdf>

FEUP at TREC 2017 OpenSearch Track Graph-Based Models for Entity-Oriented

José Luís Devezas, Carla Teixeira Lopes, and Sérgio Nunes

Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, 2017, URL: <https://trec.nist.gov/pubs/trec26/papers/FEUP-O.pdf>

Information Extraction for Event Ranking

José Devezas and Sérgio Nunes

6th Symposium on Languages, Applications and Technologies (SLATE 2017), 2017, Dagstuhl, Germany, DOI: 10.4230/OASlcs.SLATE.2017.18

Benchmarking Named Entity Recognition Tools for Portuguese

André Pires, José Devezas, and Sérgio Nunes

Proceedings of Simpósio de Informática (INForum 2017), 2017, Aveiro, Portugal

Index-Based Semantic Tagging for Efficient Query Interpretation

José Luís Devezas and Sérgio Nunes

Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings, 2016, DOI: 10.1007/978-3-319-44564-9_17

Exploring a Large News Collection Using Visualization Tools

Tiago Devezas, José Devezas, and Sérgio Nunes

Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016, 2016, URL: <http://ceur-ws.org/Vol-1568/paper9.pdf>

Large-scale crossmedia retrieval for playlist generation and song discovery

Filipe Coelho, José Luís Devezas, and Cristina Ribeiro

Open research Areas in Information Retrieval, OAIR '13, Lisbon, Portugal, May 15-17, 2013, 2013, URL: <https://dl.acm.org/doi/10.5555/2491748.2491764>

Using the overlapping community structure of a network of tags to improve text clustering

Nuno Cravino, José Luís Devezas, and Álvaro Figueira

23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25-28, 2012, 2012, DOI: 10.1145/2309996.2310036

Studying a Personality Coreference Network in a News Stories Photo Collection

José Luís Devezas, Filipe Coelho, Sérgio Nunes, and Cristina Ribeiro

Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings, 2012, DOI: 10.1007/978-3-642-28997-2_47

Creating News Context From a Folksonomy of Web Clipping

José Devezas, Henrique Alves, and Álvaro Figueira

Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2012 (IMECS 2012), 2012, Hong Kong, URL: http://cracs.fc.up.pt/sites/default/files/c2012_arf_IAENG.pdf

Using the H-Index to Estimate Blog Authority

José Luís Devezas, Sérgio Nunes, and Cristina Ribeiro

Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, 2011, URL: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2829>

Interactive Visualization of a News Clips Network: A Journalistic Research and Knowledge Discovery Tool

José Devezas and Álvaro Figueira

Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012), 2011, Barcelona, Spain, URL: <https://scitepress.org/papers/2012/41087/41087.pdf>

FEUP at TREC 2010 Blog Track: Using h-index for blog ranking

José Luís Devezas, Sérgio Nunes, and Cristina Ribeiro

Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010, 2010, URL: <http://trec.nist.gov/pubs/trec19/papers/labs-sapo-up.blog.pdf>

Studying blog features over link popularity

José Luís Devezas, Cristina Ribeiro, and Sérgio Nunes

Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNAKDD 2009, Paris, France, June 28, 2009, 2010, DOI: 10.1145/1964858.1964863

technical reports

Army ANT: A Workbench for Innovation in Entity-Oriented Search - External Option: Scientific Activities – TREC Open Search

José Devezas

Research rep., 2017, URL: <https://hdl.handle.net/10216/110181>

Auditing Open Access Repositories - Free Option: Supervised Study – Digital Archives and Libraries

José Devezas

Research rep., 2017, URL: <https://hdl.handle.net/10216/104152>

Entity-Oriented Search - FEUP InfoLab Report

José Devezas

Tech. rep., 2016

An Overview of the Graph Database Paradigm - Breadcrumbs Project

José Devezas

Tech. rep., 2011

Overlapping Community Detection - Laboratório SAPO/U.Porto Report

José Devezas, Sérgio Nunes, and Cristina Ribeiro

Tech. rep., 2011

book chapters

Creating and Analysing a Social Network Built From Clips of Online News

Álvaro Figueira, José Devezas, Nuno Cravino, and Luis-Francisco Revilla

Information Systems and Technology for Organizations in a Networked Society, chap. 5, pp. 67–86, IGI Global, 2012, URL: <https://www.igi-global.com/chapter/creating-analysing-social-network-built/76532>

theses

Graph-Based Entity-Oriented Search

José Devezas

PhD thesis [To be published], 2021

Link Ecosystem of the Portuguese Blogosphere

José Devezas

Master's thesis 2010, URL: <http://hdl.handle.net/10216/58922>

datasets

Simple English Wikipedia Link Graph with Clickstream Transitions 2018-12 [dataset]

José Devezas and Sérgio Nunes

INESC TEC research data repository, 2019, DOI: [10.25747/83vk-zt74](https://doi.org/10.25747/83vk-zt74)

selected projects

Here I present some of the most representative projects of my research career, where I developed multiple competences in the broad areas of information retrieval, network science and machine learning. This serves the purpose of illustrating not only the challenges I tackled, but also the feasibility, from start to finish, of all the projects I participated in, where a working prototype was always delivered, as a live demonstration of the carried research.

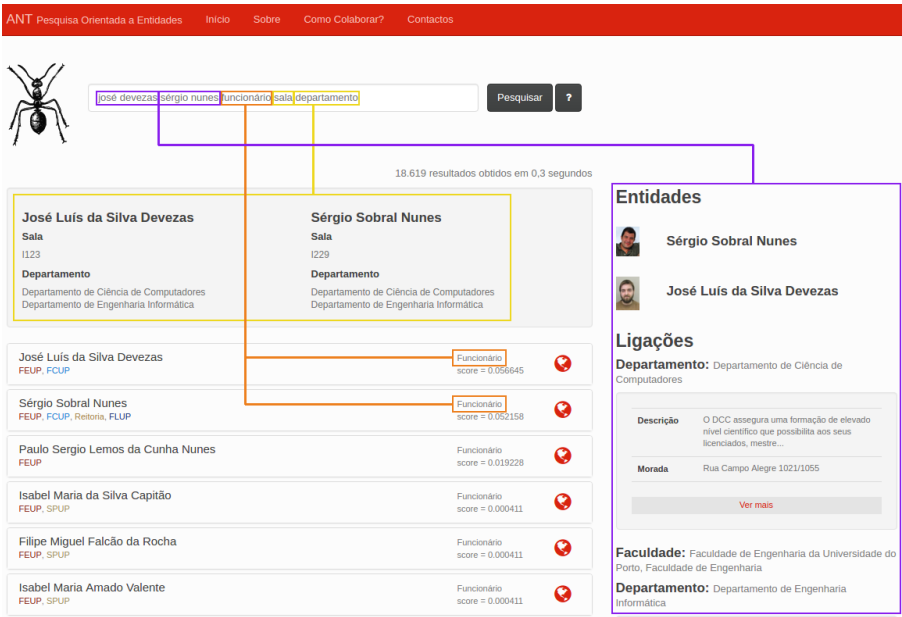
2015–2020 **ANT**

Laboratório de Sistemas de Informação/U.Porto

ANT is an entity-oriented search engine available at <http://ant.fe.up.pt>, that indexes data from the SIGARRA Information System. On one side, it aims at improving search in U.Porto. On the other side, it aims at providing a platform to educate students and to support future research in the area.

This research project motivated my doctoral proposal, providing the ideal platform for the integration of previous areas of interest, particularly information retrieval and network science, but also recommendation and personalization. Working in ANT led me to identify several challenges in the area of entity-oriented search and to devise an innovative and focused contribution.

In a similar fashion to Google and other modern search engines, this enabled our system to answer user questions more directly, well beyond the traditional keyword-based matching. Our proposed integrated solution for segmentation and annotation of queries is illustrated in the following figure of the development front-end.



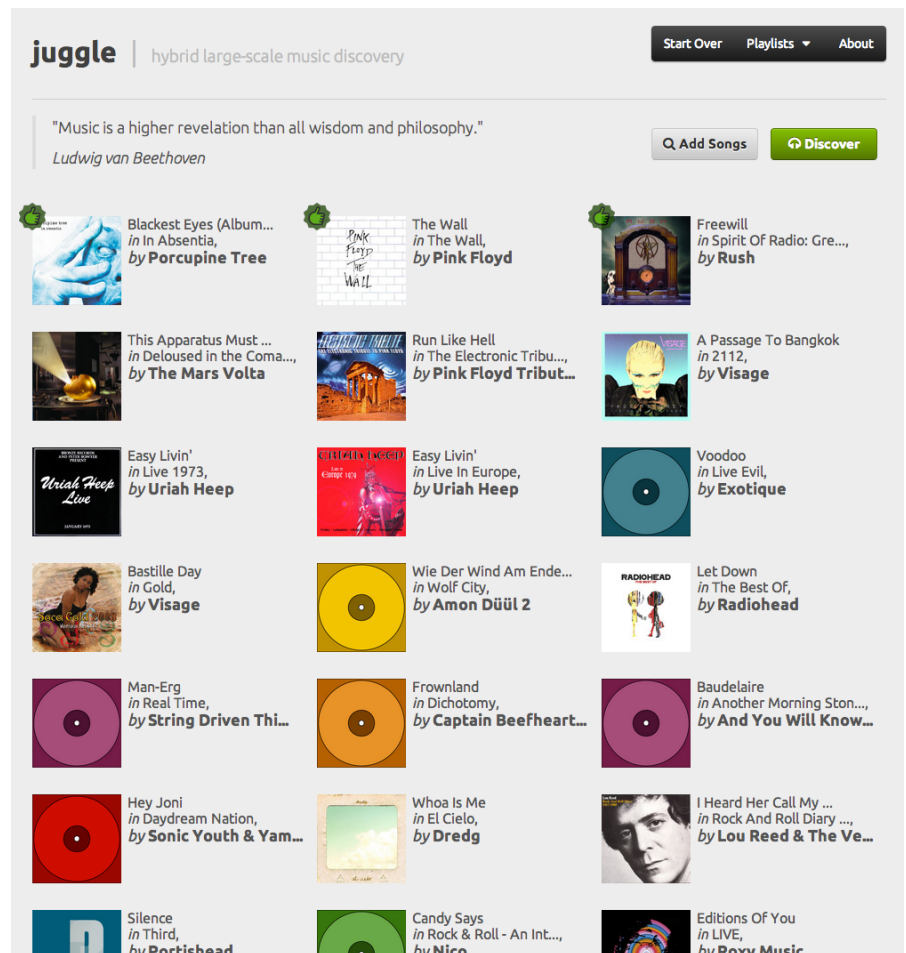
2012–2013 **Juggle**

Laboratório SAPO/U.Porto

Juggle project aimed at improving music discovery based on a hybrid large-scale recommender system, capable of handling and combining different types of data, namely text and audio content, context from elements such as tags or location, and collaborative information from user profiles.

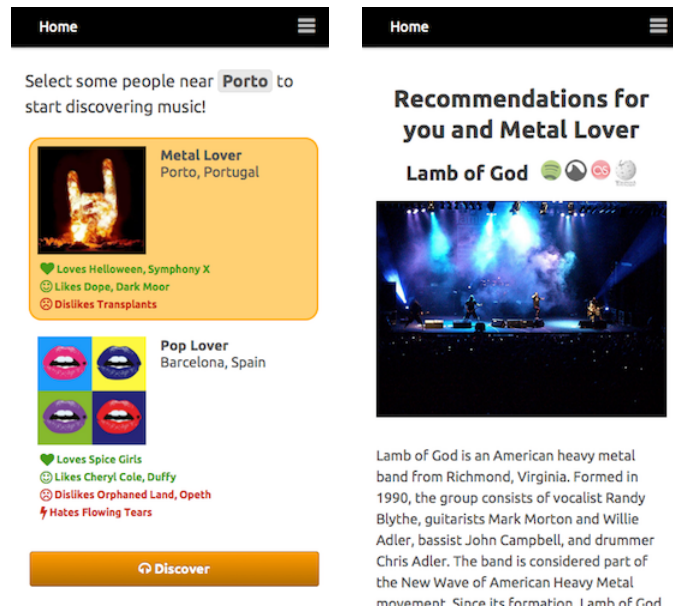
We tackled the challenges of multimodality and large-scale by developing a graph-based recommender system, supported on Neo4j, a popular and robust graph database that facilitated the modeling of content, context and collaborative information as nodes and edges in a graph. One of the biggest challenges was the translation of audio content to relationships in a graph, specifically the comparison of the audio features of a million songs with each other, which we solved by using an approximate search algorithm from image retrieval.

Our recommendation algorithm was mainly supported on neighborhood methods for collaborative filtering, but we also used metrics from text retrieval to boost the relevance of tags in the long tail, while not completely disregarding tag popularity, in order to offer a playlist that better potentiated the discovery of music.



In Juggle Mobile we presented the users with the ability to create an account and fill their taste profiles either based on our random artist rating system, or by importing their existing music information from Facebook or Last.fm. All the data from these different sources was mixed together based on our weighting model and used to provide recommendations to the user or to a group of nearby users.

Our experiments were based on a linear algebra approach, where, instead of a graph, we used a user-items matrix, applying singular value decomposition to build a latent factor model that provided the support for individual and group recommendations. For groups, we proposed a rating aggregation method that ensured an equal chance for every group member to have a relevant influence in the recommendations outcome.



2011–2012 **Breadcrumbs**

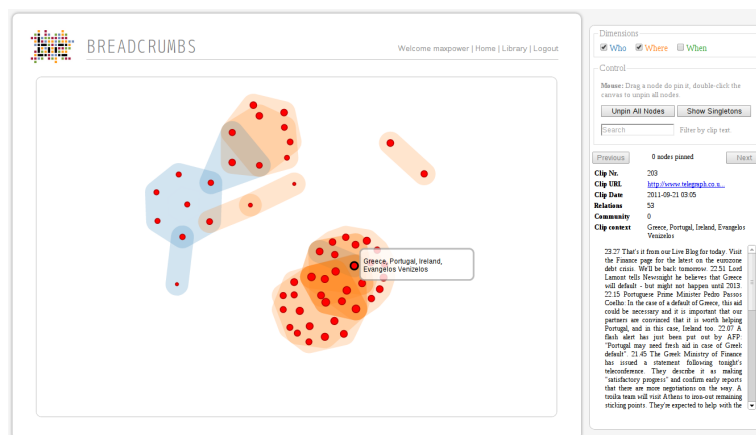
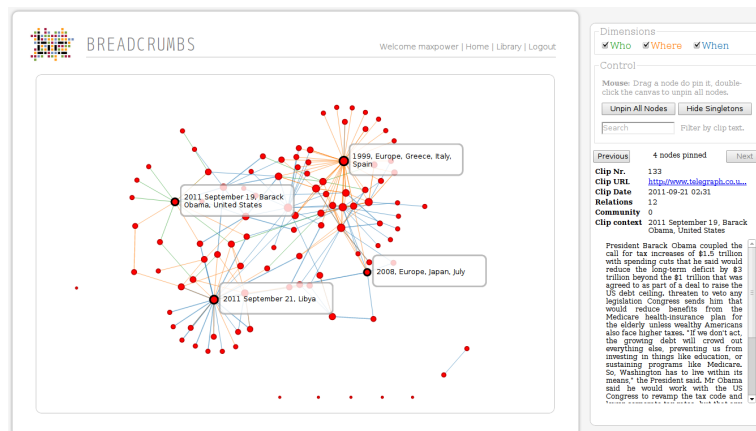
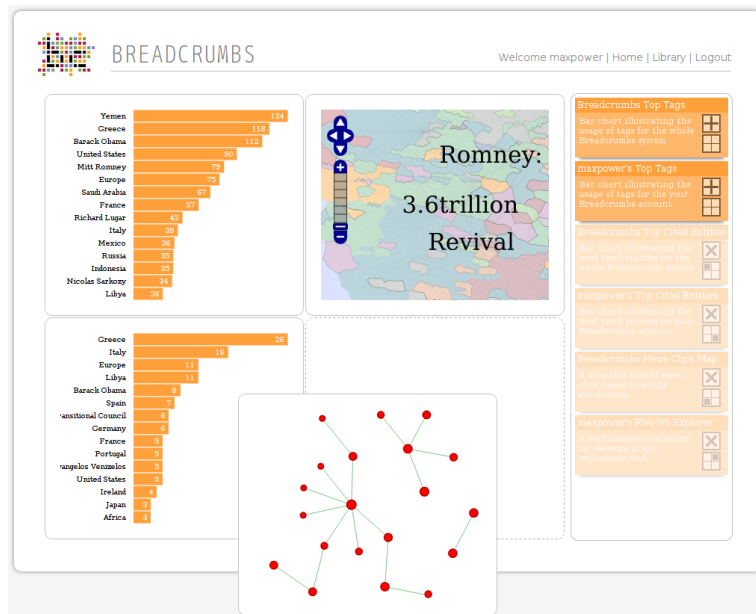
CRACS/INESC TEC

As a Breadcrumbs researcher, I was able to make contributions on several different areas. I implemented a language-independent named entity recognition system based on DBpedia entity lists. This system enabled the identification of three different types of entities — people, places and dates — tied to three of the five dimensions (the Five-Ws) of journalism: who, where and when. Using this data, a multidimensional entity coreference network was built, connecting news clips that cited the same entity. Next, I implemented the community detection methodologies for multidimensional networks proposed by Tang et al. This included the dimension integration strategies proposed by the authors, based on their unified view of four traditional community detection methodologies. These algorithms were also implemented in the system, along with the Louvain method, one of the state of the art algorithms for community detection.

Next, two visualization tools were developed to display and explore the acquired data. The first was analogous to a map, where communities were visualized as countries resulting of the aggregation of a node population. The second enabled the exploration of the multidimensional network based on the three identified dimensions: who, where and when. Some simple chart visualizations were created to display statistics about the top user and system tags and entities.

We used a topic model, based on Latent Dirichlet Allocation, to suggest titles for each collection of news clips; a simplistic event detection system was also created, in order to find relevant peaks of activity in a time series of entity frequencies. Some other trivial systems, such as an administration panel, capable of scheduling tasks, and a widget dashboard were also implemented.

These algorithms were all developed using a web services architecture, communicating using either XML or JSON. Several scientific papers were published as the results of the described research. Below are some screenshots of the Breadcrumbs modules I contributed to in some way.



2010–2011

Ciclope

Laboratório SAPO/U.Porto

One of my first projects was Ciclope, a real time data visualization project aimed at gathering information from SAPO Blogs clickstream and displaying it in a useful way, allowing the blog owner to have an understanding of how the traffic flow of his or her blog behaved.

Among other widgets, we developed two main visualizations: a real time bar chart that displayed the number of visits per second along with a table showing traffic origin and destination; and a custom flow tree to visualize and quantify aggregated traffic sources for a given blog.

