Maria Christoforaki
817 8th Avenue Apt. 4F
11215 Brooklyn, NY

(+1) 917 445 1582
mchristoforaki@gmail.com
www.cis.poly.edu/~christom/

Summary & Research Interests

I am a 5th year PhD student at NYU-Poly, working under the supervision of Professor Torsten Suel. My research interests lie at the intersection of Web Search, Data Mining, and Information Retrieval. My goal is to apply techniques of these areas in managing and analyzing large-scale datasets.

Education

- Computer Science and Engineering, Polytechnic Institute of New York University, USA,
  PhD student with GPA 4.0/4.0 (August 2009 - present)
- Computer Science and Engineering, Polytechnic Institute of New York University, USA,
  MSc with GPA 4.0/4.0 (May 2011)
- Computer Engineering and Informatics, University of Patras, Greece,
  Diploma with GPA 7.6/10 (July 2009)

Publications

- **"STEP: A Scalable Testing and Evaluation Platform"**
  M. Christoforaki and P.G. Ipeirotis
  AAAI Conference on Human Computation & Crowdsourcing, Pittsburgh, USA
  (HCOMP 2014)
- **"Estimating Pairwise Distances in Large Graphs"**
  M. Christoforaki and T. Suel
  IEEE International Conference on Big Data, Washington DC, USA
  (BigData 2014)
- **"Mining Videos from the Web for Electronic Textbooks"**
  R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan
  Intenational Conference on Formal Concept Analysis, Cluj-Napoca, Romania
  (ICFCA 2014)
- **"Text vs. Space: Efficient Geo-Search Query Processing"**
  M. Christoforaki, J. He, C. Dimopoulos, A. Markowetz, and T. Suel
  ACM Conference on Information and Knowledge Management, Glasgow, UK
  (CIKM 2011)
- **"Searching Social Updates for Topic-centric Entities"**
  M. Christoforaki, I. Erunse, and C. Yu
  International Workshop on Searching and Integrating New Web Data Sources, Seattle, WA, USA
  (VLDS 2011)

Industrial Experience

- **Google Research**, Mountain View, CA (Summer 2013)

  Software Engineering Internship: At Google Research, I interned in the NLP Research group and my work focused on the freshness of Google's Knowledge Graph (KG). In particular, I worked on the automatic extraction of fresh information from annotated news articles to be added to the KG. The KG is Google's knowledge base and it is used for enhancing its search engine's results with semantic information. During my internship I developed a workflow extracting information from news articles as well as a machine-learned model for validating this information using various related signals. Our experiments on a subset of predicates indicated that this method has very high recall while keeping precision at a satisfying level.
  Supervisor: Dr Min Wang (minwang@google.com)

- **Microsoft Research**, Mountain View, CA (Summer 2012)

  Research Internship: At Microsoft Research, I interned in the Search Labs group and my work focused on automatically augmenting educational textbooks with videos. Textbooks are generally organized into sections such that each section is focused on explaining few concepts and every concept is primarily explained in one section. Building upon these principles from the education literature, we proposed techniques for identifying the focus of a section in relationship to other sections in the book in terms of a few indicia, which themselves are unique combinations of concept phrases. We scored the candidate videos for augmenting a textbook section based on concept phrases from that section's indicia present in the auditory tract of the video. Our tests using two corpora of textbooks indicated that our system is able to find useful videos.
  Supervisor: Dr Anitha Kannan (ankannan@microsoft.com)

- **Linkedin**, Mountain View, CA (Summer 2011)

  Data Analytics Internship: At Linkedin, I interned in the Data Analytics group and my work focused on the InMaps product. InMaps provides a visualization of each member's Linkedin network (egonet). In particular I worked on the community detection algorithm that colors the egonet vertices as well as on the automatic label suggestion of the detected communities. Furthermore I worked on efficiently updating egonets and community structure when new members are added or old ones are deleted. For the implementation I used Java and pig and the final workflow was deployed on the Hadoop cluster of Linekdin. The data produced by my work is leveraged not only by InMaps but by other Linkedin applications as well.
  Supervisor: Mathieu Bastian (mbastian@linkedin.com)

- **CERN**, Geneva, Switzerland (Summer 2007)

  OpenLab summer student programme: Participated in a project for an efficient simulation of the BitTorrent protocol and of chocking and unchoking algorithms, by extending existing network simulator software. This demanded an in-depth understanding of the structure and advantages of PlanetSim (implemented in Java), the simulator we used, and the BitTorrent protocol. The implementation included adding new classes in the application layer and new scripts for the simulation to run.
  Supervisor: Dr Magdalena Punceva (magdalena.punceva@epfl.ch)

- **CERN**, Geneva, Switzerland (Summer 2006)

  Internship: In cooperation with the CMS-experiment group, I designed and implemented a web-based application for conferences concerning the CMS-experiment and their candidates, using a multi-tier architecture based on MySQL, Java Servlets, JSP and XML technologies. Among other services, the system also provided candidate-application management and generation of conference activity statistics for the conference team to interpret and handle.
  Supervisor: Prof Maria Spiropulu (smaria@cern.ch)

Ongoing Research

- **Learning Term Impact Scores of Complex Rankers for Efficient Query Processing.**
  We are studying the trade-off between efficiency and result quality for modern search systems. Such systems commonly use 2-tier ranking schemes during query processing: the first tier uses a simple but fast ranking function to perform a top-k candidate selection; the second tier then refines these results with a complicated and slower machine learned ranking function. We propose a first tier ranking function that is learned via the output of the second tier one, which reduces the number of documents passed to the second tier while retaining high result quality.
  (with Constantinos Dimopoulos and Torsten Suel)

Technical Skills

- Programming Languages: Java, C, C++, C#, Python, R
- Data management: Hadoop (pig and MapReduce), SQL

Selected Coursework

Advanced Database Systems, Web Search Engines, Advanced Algorithms and Data Structures, Computational Geometry, Machine Learning, Artificial Intelligence, Distributed Systems, Operating Systems, Computer Networking, Software Engineering

References

- Prof. Torsten Suel (suel@poly.edu)
  Polytechnic Institute of New York University
- Prof. Panos Ipeirotis (panos@stern.nyu.edu)
  New York University
- Dr. Anitha Kannan (ankannan@microsoft.com)
  Microsoft Research
- Dr. Cong Yu (congy@umich.edu)
  Google Research