O Scientist, Where Art Thou?
Affiliation Propagation for Geo-Referencing Scientific Publications

Babak Ahmadi and **Salah Zayakh** and **Fabian Hadiji** and **Kristian Kersting**
Knowledge Dicovery, Fraunhofer IAIS, 53754 Sankt Augustin, Germany
{firstname.lastname}@iais.fraunhofer.de

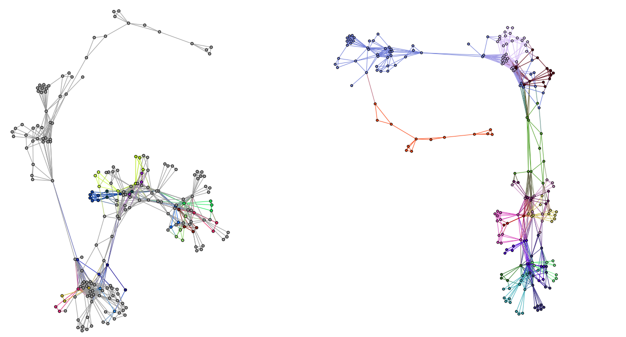
Abstract

Today, electronic scholarly articles are available freely at the point of use. Moreover, bibliographic systems such as DBLP, ACM's Digital Libraries, Google's Scholar, and Microsoft's AcademicSearch provide means to search and analyze bibliographic information. However, one important information is typically incomplete, wrong, or even missing: the affiliation of authors. This type of information can be valuable not only for finding and tracking scientists using map interfaces but also for automatic detection of conflict of interests and, in aggregate form, for helping to understand topics and trends in science at global scale. In this work-in-progress report, we consider the problem of retrieving affiliations from few observed affiliations only. Specifically, we crawl ACM's Digital Libraries for affiliations of authors listed in DBLP. Then, we employ multi-label propagation to propagate the few observed affiliations through out a network induced by a Markov logic network on DBLP entries. We use the propagated affiliations to create a visualization tool, PubMap, that can help expose the affiliations, using a map interface to display the propagated affiliations. Furthermore, we motivate how the information about affiliations can be used in publication summarization.

1 Introduction

It's late September 2011. You have just returned from Magdeburg, where you attended KDML 2011. You turn on your entertainment center at home and find your photos from the trip already there; uploaded through a combination of Wi-Fi and cellular connections your camera detected along the way. On one of the photos you see Stefan Wrobel, the speaker of GI Special Interest Group KDML. "Ah, I really enjoyed talking to him", you reminisce as you flip through the photos. You wonder what he is currently working on. You speak into the EC microphone: "Show me the career of Stefan Wrobel". The EC delivers Stefan's career in a format that's quintessentially human, namely via storytelling. It shows a slide show of Stefan's geo-located affiliations and major publications with a voice-over telling. The cyberdog finally arrives with your Koelsch. What a wonderful life. This vision may become a reality well after 2011, but it is actually not far-fetched.

The world-wide web contains an overwhelming amount of information which is growing dramatically. Search engines enable us to efficiently find information that is likely to be most relevant for our query and return the results in a list containing the top k hits. The exponential growth of data is indeed challenging. However, it also gives opportunities for many novel applications that have hardly been possible before. We can ask for example whether a machine-readable list is the best way to present such data or can we improve the way information is presented to the user? Consider for example the computer science bibliography DBLP [1]. It is a rich source of information. But most of the information is hardly used and the results of single queries is only presented in a list to the user. DBLP stores the scientific publications of authors but also lots of inherently relational information like co-authorship of papers for example. Would it not be great if we could search for an author and not only get her papers and co-authors listed, but, for example, also display the affiliations a scientist was with in her career. In this paper we describe ongoing work that tackles this problem. We obtained the list of publications for each author from DBLP. The publication lists do not contain any information about the affiliation of the authors. These can partially be crawled from the ACM Digital Library [2]. The parsed information, however, is incomplete. For a large amount of papers the affiliations are not available. So we need to find a way to fill in the gaps.



(a) Labels parsed from ACM-DL. Each color represents a different affiliation. Only few nodes are colored (labeled), others are grey (unlabeled).

(b) Labels after propagation: The labels have propagated through the entire network. Previously unlabeled nodes are now colored.

Figure 1: Affiliation graph

Therefore, we describe a propagation logic to set the graph structure and to assign appropriate weights to the edges (Sec.2). On this graph we distribute labels for un-

[1] http://dblp.uni-trier.de/

[2] http://portal.acm.org/

known data points using label propagation (Sec.3). In Section 4 we describe the task of publication summarization and indicate how an author's affiliations can enhance the summarization quality. Throughout the work we show preliminary results of the current work in progress.

2 Propagation Logic

We can use the underlying database to construct our graph that we use to propagate the affiliations. Therefore we construct an adjacency matrix A with a row and a column for each pair of paper and its authors in our database. We define a Markov Logic like model to set the weights of our matrix as follows. We have facts over:

- $\texttt{year(Y)}$,
- $\texttt{author(A)}$,
- $\texttt{paper(P)}$,
- $\texttt{publicationYear(P,Y)}$,
- $\texttt{hasAuthored(A,P)}$,
- and $\texttt{succDBLP}(\texttt{P}_1,\texttt{P}_2)$.

The additional predicate $\texttt{succDBLP}(\texttt{P}_1,\texttt{P}_2)$ encodes the total order among paper of one order given by DBLP. More importantly, after initializing A to $\mathbf{0}$, where $\mathbf{0}$ is an all-zero matrix we set the weights of A according to rules such as

$$\begin{aligned} A(\texttt{I,P;J,P}) += \; w_1 \;\; &\textbf{IF}\; \texttt{hasAuthored(I,P)} \\ &\wedge\; \texttt{hasAuthored(J,P)} \end{aligned}$$

This adds an edge (a_{ip}, a_{jp}) for two co-authors i and j of the paper p. If the edge already exist, the edge weight is incremented by w_1.

$$\begin{aligned} A(\texttt{I},\texttt{P}_1;\texttt{I},\texttt{P}_2) += \; w_2 \;\; &\textbf{IF}\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_1) \\ &\wedge\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_2) \\ &\wedge\; \texttt{publicationYear}(\texttt{P}_1,\texttt{Y}) \\ &\wedge\; \texttt{publicationYear}(\texttt{P}_2,\texttt{Y}) \end{aligned}$$

This adds an edge (a_{ip_1}, a_{ip_2}) (or adds w_2) if two papers of the same author have been published in the same year.

$$\begin{aligned} A(\texttt{I},\texttt{P}_1;\texttt{I},\texttt{P}_2) += \; w_3 \;\; &\textbf{IF}\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_1) \\ &\wedge\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_2) \\ &\wedge\; \texttt{publicationYear}(\texttt{P}_1,\texttt{Y}) \\ &\wedge\; \texttt{publicationYear}(\texttt{P}_2,\texttt{Y}+1) \end{aligned}$$

This adds an edge (a_{ip_1}, a_{ip_2}) (or adds w_3) if two papers of the same author have been published in subsequent years.

$$\begin{aligned} A(\texttt{I},\texttt{P}_1;\texttt{I},\texttt{P}_2) += \; w_4 \;\; &\textbf{IF}\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_1) \\ &\wedge\; \texttt{hasAuthored}(\texttt{I},\texttt{P}_2) \\ &\wedge\; \texttt{succDBLP}(\texttt{P}_1,\texttt{P}_2) \end{aligned}$$

This adds an edge (a_{ip_1}, a_{ip_2}) (or adds w_4) if one paper is the successor of another paper of the same author in the ordering given by DBLP.

The weights w_1, …, w_4 are set additively and reflect the relative importance of each rule. These do not necessarily imply that the affiliation has to be propagated, e.g. from one author to the next. They simply express that it is more likely that one person P has affiliation I if a certain rule applies. Figure 1 **(a)** shows for example the graph constructed for Stefan Wrobel. Here we can see the labels that were crawled from the ACM Digital Library. Each color represents a different affiliation. As one can see only few nodes are colored (labeled) and due to incomplete data most nodes are grey (unlabeled). This graph is then used to propagate the labels to previously unknown papers as will be shown in the next section.

3 Affiliation Propagation

Now to propagate the known affiliations in the graph such that we obtain labels for nodes of the graph where the label is unknown, we can employ a random walk with restart (RWR) technique on the graph induced by the adjacency matrix A, called label propagation (LP), see e.g. [Zhou *et al.*, 2003] and [Bengio *et al.*, 2006].

In label propagation we use the known labeling of our nodes and the graph structure which is given by our adjacency matrix to propagate whether the neighbouring nodes should also have a given label or not.

More formally, we compute the diagonal degree matrix $\mathbf{D}$

$$\mathbf{D}_{ii} = \sum\nolimits_j \mathbf{A}_{ij}.$$

Then, we compute the normalized graph Laplacian

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2},$$

and initialize a column vector $\mathbf{y}$ according to the labels we have, that is $\mathbf{y}_i = 1$ if node i is labeled positive, $\mathbf{y}_i = 0$ otherwise. The final labeling $\mathbf{x}$ can then be computed by solving the following system of linear equations

$$(\mathbf{I} - \alpha\mathbf{L})\mathbf{x} = (1-\alpha)\mathbf{y}.$$

An entry x_i in the result-vector x intuitively can be interpreted as the likely label of node i. If x_i is close to one it is labeled positive.

In our setting, however, each possible affiliation of an author and her co-authors is a label and thus we have to solve the problem of multi-label propagation. Given a point set of $X = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\}$ and a label set $L = \{1, \ldots, c\}$ the first l points $x_1, \ldots, x_l$ are labeled as $y_i \in L$ and the remaining are unlabeled. We can, however employ a similar strategy as in the single-label case but instead of having a single label vector, we have one vector for every label $l \in L$. Now we have to solve multiple systems of linear equations:

$$(\mathbf{I} - \alpha\mathbf{L})\mathbf{X} = (1-\alpha)\mathbf{Y}.$$

The label of each point x_i is given by

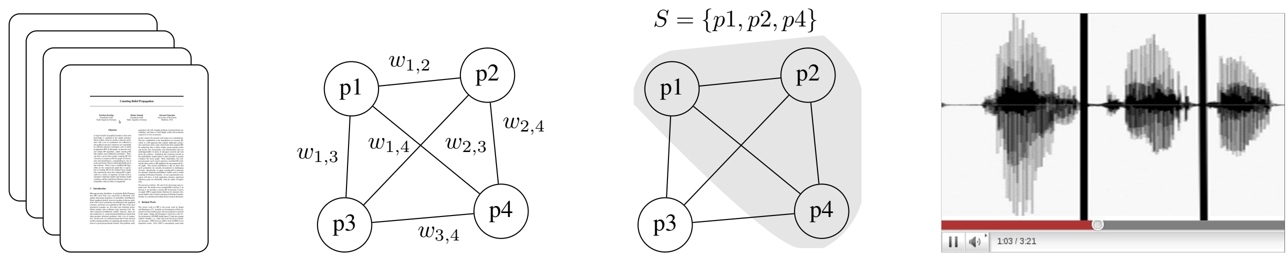
$$y_i = \arg\max_{j \le c} Y(:, i),$$

where $Y(:, i)$ denotes the i-th row of the matrix Y.

Using affiliation propagation we now obtain labels for papers that we had no information about. Figure 1 **(b)** shows the graph after the labels have propagated through the entire network. Previously unlabeled nodes now have been assigned a color and can be visualized on the map.[3]

Figures 3 and 4 show the courses of the careers of Stefan Wrobel and Luc De Raedt respectively. In our system we have a slider to move from the very first paper of an author to the most recent papers (not shown in the figures). The affiliation of the papers that are currently viewed are marked on the map, i.e. if all papers current have the same affiliation the dots coincide and we see only one marked location on the map. In Figures 3**(a)-(c)** one can see that according to the system Stefan Wrobel has started his career in Berlin. Then he moved to Magdeburg later to

[3] Note that, although Stefan Wrobel had only four different affiliations the graph contains more colors. This is due to the fact that there is often the same affiliation spelled differently or abreviated. Thus, one should also enhance the overall system by affiliation resolution also for names of affiliations that are highly ambiguous. This is part of our future work.



(a) **Step 1:** Get papers $P = \{p_1, \ldots, p_n\}$ for an author.

(b) **Step 2**: Define similarity $w_{i,j}$ between papers and determine costs c_i for each.

(c) **Step 3**: Use greedy algorithm to select subset S based on $w_{i,j}$, c_i, and B.

(d) **Step 4**: Generate the final audio clip based on papers in S.

Figure 2: Steps in publication summary.

Bonn/St.Augustin. In reality, however, Stefan Wrobel has already been in Bonn once before he moved to Magdeburg. As there was no evidence for the first stay in Bonn unfortunately this could not be inferred by the system. We are currently working on simultaneous propagation for multiple authors, ideally all authors contained in DBLP, to let them influence each other. For example, a supervisor has a lot of influence on the affiliation of a young researcher and should be able to complete missing information as they have usually a lot of papers in common. We have initial very promising results on performing affiliation propagation on the joint graph of multiple authors. Figures 4**(a)-(c)** show that the affiliation of Luc De Raedt have correctly been propagated. He has started in Leuven, then moved to Freiburg and back to Leuven. In Figure 4**(b)** the papers that are currently viewed have two different affiliations. Thus we also have multiple affiliations marked on the map.

4 Publication summarization

Reconsidering our example from the introduction, now we would like to get an overview of the career of Stefan Wrobel without searching the entire web to collect every single piece of information ourself and we want to go beyond the information that is available on his current webpage. One can easily think of scenarios in which an author's webpage is not informative enough or a different way of representation is more entertaining to the user. Besides a list of his affiliations, it is also interesting to briefly go through a list of his most influential and representative publications. Optimally, this kind of short CV of an author should be presented as a multimedia clip, containing audio, images, videos, and possibly more. You can personalize the clip by choosing the maximal length of the clip and possibly further properties such as a focus on a timespan. Based on the chosen duration, a short clip is generated which presents valuable information on the author based on the data gathered from the web.

One essential part of such a short CV would be a well selected list of publications. Young scientists might only have a few publications in a single field therefore the selection could be based mainly on the quality of the corresponding venues. On the other hand, scientists which publish in their research field for years or even decades have published on several different topics together with numerous different co-authors. In these cases, several further features can be used in the selection process and a change of the affiliation might indicate a shift in topics as well.

As a first step towards a multimedia clip, we start by generating the short CV based solely on the scientist's publication history in combination with the history of her affiliations and citation counts of the papers. We obtain the publications history P for an specific author from DBLP (Fig. 2(a)) and cast the problem into a text summarization framework to select an expressive subset. In comparison to text summarization, here, the full list of publications corresponds to a text document and a single paper relates to a single sentence in the document. Based on this representation we want to select a subset S of the papers which describes the entire set best, and forms the basis for our short CV. In text summarization, the selection is usually based on a bag-of-words model while we focus here on the metadata of papers. Additionally, we want to incorporate information such as the citation count which should influence the selection on top of the pairwise similarity of two papers.

We adapt the text summarization approach presented in [Lin and Bilmes, 2010] for our publication selection problem. The work presents a greedy algorithm which is based on a submodular quality function. This quality function is maximized under a budget constraint by iteratively adding sentences to the summary as long as the quality of the summary is improved and the given budget is not exceeded. The submodularity of the quality function ensures that the greedy algorithm selects a near-optimal solution. The quality function is defined on the current summary, i.e. a subset of sentences from the document, and takes the pairwise similarity of sentences into account.

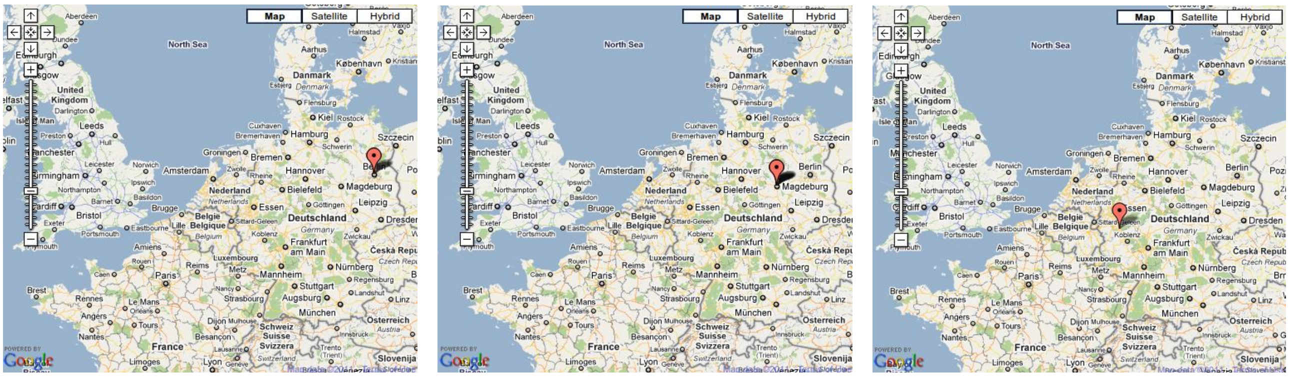
We adapt this approach by defining a similarity measure $w_{i,j}$ on publication entries (Fig. 2(b)). To determine the costs c_i of a paper, we phrase every publication by a predefined template sentence and use a speech synthesis module to convert it to audio. There are various template sentences defined in advance and for each publication a suitable one is selected in the summarization process. Our budget constraint B is the user defined length of the audio clip.

One example of a template sentence might look like the following:

While visiting **W** *in* **X**, **Y** *published the paper* **Z**.

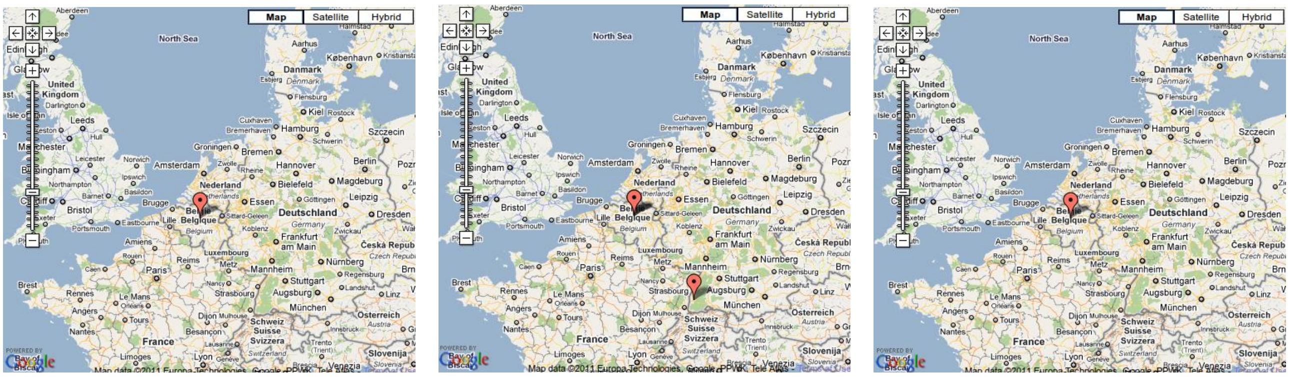
Here, **W** will be replaced by the author's affiliation in year **X** while **Y** and **Z** correspond to the author's name and the title of the paper. Depending on the length of the title and the other placeholders, the duration of resulting audio clip will vary.

We apply the greedy algorithm to the publication history of an author (Fig. 2(c)) and select a subset so that the total length of the concatenated audio files is below a user chosen duration. Based on the chosen papers we can then create the final audio clip from the single clips (Fig. 2(d)).



(a) Berlin (b) Magdeburg (c) Bonn

Figure 3: The course of affiliations for Stefan Wrobel. Stefan Wrobel has started his career in Berlin. Then he moved to Bonn, Magdeburg and back to Bonn/St.Augustin. The affiliations have been propagated. However, as there was no evidence for the first stay in Bonn this could not be visualized.



(a) Leuven (b) Transition from Freiburg to Leuven (c) Leuven

Figure 4: The different affiliations at the different stages of the career of Luc de Raedt starting in Leuven, then moving to Freiburg and after some years back to Leuven. The affiliations have been correctly identified and are displayed above.

5 Conclusion and Future Work

In this work-in-progress report, we have introduced affiliation propagation. We have shown how a small number of axioms in a Markov logic like language capture the essential features of a multi-label propagation approach to populate DBLP with affiliations and the corresponding geo-locations.

There are many attractive avenues for future work. By analyzing the affiliations of publications, one can generate tag maps that mirror Milgram's well-known manually created attraction map. Intuitively, we extract topics of documents located in the correct view using e.g. latent Dirichlet allocation. Then, we overlay the topics on the map in their "capture" location. This way, we get an easy-to-grasp view on, say, the research topics in Germany. Or, we can easily answer the questions "What do European researchers work on?" What are US American researchers interested in? Another avenue is to use the affiliations to find a region's persons who are knowledgeable about a given topic. For instance, a German company may need an expert on kernel-machines in Hamburg, i.e., Who are the experts in kernel machines near Hamburg? In any case, one should also enhance the overall system by affiliation resolution. Names of affiliations are highly ambiguous. In particular, it can happen that a name of an affiliation refers to multiple places. Consider e.g. the affiliation DFKI. There are DFKIs in Saarbrücken and Bremen. Many affiliations are also referenced by multiple names, e.g. MIT vs. Massachusetts Institute of Technology or simply due to typos. This is an instance of the general problem of entity resolution, the problem of determining which records in a database refer to the same entities, and is a crucial and expensive step in the data mining process. Here, statistical relational models have been proven highly successful. So far, however, they have mainly considered discrete information such as names. We are currently working on statistical relational approaches that can deal with continuous values such as the geo-locations extracted in the present paper.

References

[Bengio *et al.*, 2006] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, pages 193–216. MIT Press, 2006.

[Lin and Bilmes, 2010] Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *NAACL/HLT*, 2010.

[Zhou *et al.*, 2003] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with Local and Global Consistency. In *NIPS*, 2003.