Tochtermann, Maurer (Eds.): Proceedings of the I-KNOW ’08, Graz
8th International Conference on Knowledge Management
Journal of Universal Computer Science, pp. 358-365, ISSN 0948-695x

Weighted Experts: A Solution for the Spock Data Mining Challenge

Benno Stein and **Sven Meyer zu Eissen**
(Faculty of Media / Media Systems
Bauhaus University Weimar, Germany
{benno.stein|sven.meyer-zu-eissen}@medien.uni-weimar.de)

Abstract: One of the most popular and trend-setting Internet applications is People Search on the World Wide Web. In its most general form, information extraction for persons from unstructured data is extremely challenging, and, we are pretty far away from satisfying solutions. However, current retrieval technology is able to cope with restricted variants of the problem, and this paper deals with such a variant, the so-called *multi document person resolution*. Given is a set of Web documents, and the task is to state for each document pair whether the two documents are talking about the same person or not.
For this problem Spock Inc., Silicon Valley, launched in 2007 a competition offering a grand prize of $50 000. Task was the person-specific classification of 100 000 Web pages within 4 hours on a standard PC, striving for a maximum F-Measure. The paper in hand describes the challenge and introduces the technology of the winning team from the Bauhaus University Weimar [see 1].

Key Words: person resolution, named entity recognition, supervised cluster analysis
Category: H.3.1, H.3.2, I.5.3, M.7

1 Person Resolution

If one were able to assort for an arbitrary person x the entire information that people (including the person him/herself) contributed about x on the Web, a database of enormous value could be compiled. The benefits of such a database reach from private interests, e.g., when looking for ancient classmates, up to expert search, hiring services, or person-specific advertisement. Under `www.search-engine-index.co.uk/People_Search`, for example, various people search services can be found, and, the list is by far not complete. Typically, such services provide some kind of full-text search, possibly restricted to certain categories, person characteristics, or other constraints. If the underlying database is maintained by human editors the quality of the search results will be very high. However, since unstructured data is the fastest growing information source nowadays, there is the question whether such repositories can be built up automatically, or whether the desired information can be extracted in an ad-hoc manner.

Consider in this connection a person query (= the name of a person enclosed in quotation marks) entered into the Google interface, which ideally should yield an assorted result list, gathering all documents of the same person into its own class. In practice the mapping between people and their names is not one-to-one, and hence the search result contains Web pages of different individuals having the same name. Moreover, since an individual can have several Web pages, search results get cluttered, especially when searching for people with common names. The outlined problem can only be addressed with a deeper semantic analysis: Web page content must be interpreted in multiple respects in order to distinguish between different individuals, even if they have the same name. This grouping problem is referred to as “multi document person resolution” [Fleischman and Hovy 2004]; it has recently gained much attention, among others through the Spock Data Mining Challenge.

[1] S. Meyer zu Eissen, B. Stein, St. Becker, Ch. Bräutigam, T. Rüb, and H.-C. Tönnies.

1.1 The Spock Data Mining Challenge

> "A common problem that we face is that there are many people with the same name. Given that, how do we distinguish a document about Michael Jackson the singer from Michael Jackson the football player? With billions of documents and people on the web, we need to identify and cluster web documents accurately to the people they are related to. Mapping these named entities from documents to the correct person is the essence of the Spock Challenge."
>
> [http://challenge.spock.com]

Definition 1 Multi Document Person Resolution Problem. Given are a set of (Web) documents $D = \{d_1, \ldots, d_n\}$, a set of referents $R = \{r_1, \ldots, r_m\}$, a set of person names $Ne = \{ne_1, \ldots, ne_l\}$, and a set of target names Nt, $Nt \subset Ne$. Moreover, let $\nu(d) : D \to \mathcal{P}(Ne)$ designate a function that returns all person names contained in a document d. Based on these sets the tuple $\langle D, R, Ne, Nt, \beta \rangle$ defines a multi document person resolution problem, π_{pr}, if the following conditions hold:

(1) $\beta : R \to Nt$ is a mapping that assigns a target name to each referent.

(2) $\forall_{d \in D} : |\nu(d) \cap Nt| = 1$

Finally, we call each function $\gamma : D \to R$ a solution of π_{pr}.

Remarks. (*i*) The set of referents R is the complete set of interesting persons the documents in D talk about. Related to the quotation above the name Michael Jackson is a target name, and Michael Jackson the singer as well as Michael Jackson the football player are referents. (*ii*) The fact of being a function qualifies γ as a solution for π_{pr}. Note, however, that a reasonable solution of π_{pr} will fulfill $\beta(\gamma(d)) \in \nu(d)$ for all $d \in D$; i.e., the referent's name occurs in its associated documents. (*iii*) In order to evaluate a solution γ of π_{pr}, a reference classification is needed; the quality of γ can be expressed in terms of the F-Measure, for example. (*iv*) A document d may contain several person names $\nu(d)$ from which exactly one must be a target name. Of course, the multi document person resolution problem can be defined differently; e.g., one could allow more than one target name per document. The above definition reflects the constraints of the Spock Data Mining Challenge.

Figure 1 shows the benchmark data of the challenge. The training data is a person resolution problem π_{pr} for which the correct classification γ^* is given; γ^* is also designated as ground truth (of the training data) and has an F-Measure value of 1 per definition. The data set was built by Spock using their Web crawler; observe the imbalanced distributions of the number of referents per target name and the number of documents per referent. The challenge started in April 2007 and ran till December 2007.

1.2 Existing Research

Entity resolution exists in several variants, imposing constraints on whether or not training data is available, on whether or not external information sources can be used, or on the maximum time for computing the entity clusters. Some approaches are general enough to work across domains, e.g. the research described in [Pasula et al. 2002;

Number of		Gbyte
target names	44	
referents	1 101	
training documents	27 000	2.3
test documents	75 000	7.8

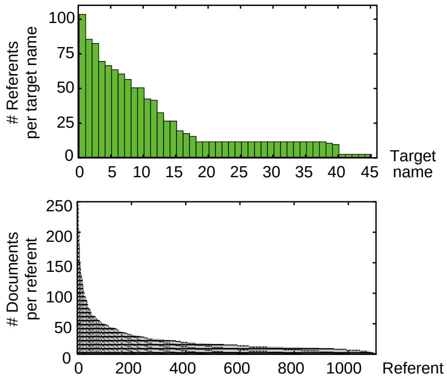

Figure 1: Benchmark data of the Spock Data Mining Challenge (left), the distribution of the referent number over target names (right top), and the distribution of the document number over referents (right bottom).

Daumé III and Marcu 2005; Bhattacharya and Getoor 2006]. The latter two approaches employ latent Dirichlet models for entity resolution, which have been proven to perform well for author name disambiguation, for proper noun coreferencing, and for reference matching. However, these approaches have not been adapted or analyzed for the assorting of Web pages according to referents. [Mann and Yarowsky 2003] propose a clustering approach to disambiguate documents containing target names. Their underlying retrieval models employ term weight features, similarity features based on proper nouns, and similarity features emerging from extracted biographic data like birth year, occupation, and school. They test their approach on a small set of Web pages, which were retrieved from target name queries to Google and manually sorted according to referents. The performance is about 86% accuracy. [Fleischman and Hovy 2004] defined the "multi document person name resolution" problem for which we gave a formalization in Definition 1. The authors propose a set of features to construct a binary classifier that decides whether two documents talk about the same person; the feature set comprises external knowledge acquired from Web search engines. The results of the classification are translated into a similarity graph which then is clustered. The authors experiment with a set of Web pages referring to 31 target names from the ACL dataset. Their results are difficult to interpret since the underlying performance metric is unclear. [Bekkerman and McCallum 2005] use agglomerative clustering to identify referents in Web pages. For evaluation purposes the authors collected Google search results for 12 target names. After result cleaning, the data comprised 1085 Web pages referring to 187 referents. For this small dataset the authors report an F-Measure of 0.80 with their approach.

From its form the last setting resembles the Spock Data Mining Challenge: Web pages are crawled and made accessible offline, training data is available, and no online requests to Web services are permitted to classify the test data. This scenario is realistic since a service for person name resolution needs to build up a tailored index of relevant pages, avoiding to spend money for frequent search engine queries. Note, however, that the above approaches were evaluated for small and manually cleaned datasets only; it is unclear whether they scale-up in terms of runtime performance and accuracy when given datasets of realistic size. Especially a manual data cleaning is unrealistic when dealing with Gbyte orders of magnitude.

1.3 Contributions

Key contribution is the development and implementation of technology to compute a solution γ for the person resolution problem π_{pr} specified in Figure 1, reaching an F_α-value of 0.40 with $\alpha = 1/3$ [see 2]. Section 2 introduces the main building blocks of our technology, Section 2.2 reports on selected analysis results, and Section 3 concludes by pointing to different places for improvement. Our solution includes new retrieval models, new ways to combine retrieval models with classification and data mining technology, and deals with realistic orders of magnitudes.

2 Elements of Our Analysis Technology

In an open and dynamic environment like the World Wide Web the number of referents $|R|$ is high and subject to frequent changes. This means that π_{pr} cannot be tackled by a supervised multi-class, single-label classification approach, where the referents R define the classification scheme according to which the documents D are classified. Instead, π_{pr} must be understood as a clustering problem: objective is the formation of maximum groups each of which is associated with a single referent. However, if a representative sample of documents along with its correct classification γ^* is given, one can learn from this data, e.g. particular parameters for the clustering algorithm, which in turn can be applied for the clustering of unseen samples. Due to their hybrid nature, clustering algorithms that use knowledge from training data are called “supervised clustering algorithms” [Daumé III and Marcu 2005; Finley and Joachims 2005].

Our approach combines o retrieval models, $\mathcal{R}_1, \ldots, \mathcal{R}_o$, to capture the similarity between two documents as an o-dimensional *meta similarity* vector $\varSigma$. In addition, we apply the supervised cluster analysis paradigm: based on the set of meta similarity vectors $\mathbf{\Sigma}_{train}$ along with the correct classification γ^*_{train} for a sample D_{train}, a document pair classifier $\delta : \mathbf{\Sigma} \rightarrow [0;1]$ is learned (see Figure 2). Ideally, δ will assign to the meta similarity vector $\varSigma(d_1, d_2)$ a value close to zero if d_1 and d_2 are talking about two different referents—and a value close to one otherwise. In this sense, $\delta(\varSigma(d_1, d_2))$ can be considered as an estimate for the similarity between the documents d_1 and d_2; Figure 3 illustrates for this similarity the desired probability distributions to which a combination of ideal retrieval models and classification technology should adhere.

The steps of the overall analysis process are listed below and illustrated in Figure 2. The process comprises a pre-processing stage for the construction of δ, and an application stage, where δ is used to compute a similarity graph G_δ for the test data, which then is merged.

Classifier Construction:

(1) $\forall d \in D_{train}$: compute the retrieval model representations $\mathbf{d}_\mathcal{R}$, $\mathcal{R} \in \{\mathcal{R}_1, \ldots, \mathcal{R}_o\}$.

(2) $\forall d_1 \forall d_2 \in D_{train}$: compute the model-specific similarity values $\varphi_\mathcal{R}(\mathbf{d}_{1,\mathcal{R}}, \mathbf{d}_{2,\mathcal{R}})$, $\mathcal{R} \in \{\mathcal{R}_1, \ldots, \mathcal{R}_o\}$. The meta similarity vector $\varSigma(d_1, d_2)$, $|\varSigma| = o$, comprises the similarities for (d_1, d_2); $\mathbf{\Sigma}_{train}$ is the set of meta similarity vectors for D_{train}.

[2] The $prec$- and rec-values are computed from the fraction of correctly classified document pairs in relation to all document pairs. In particular, $F_\alpha = (\alpha+1)/(\frac{\alpha}{prec} + \frac{1}{rec})$, which puts extra emphasis on the precision.

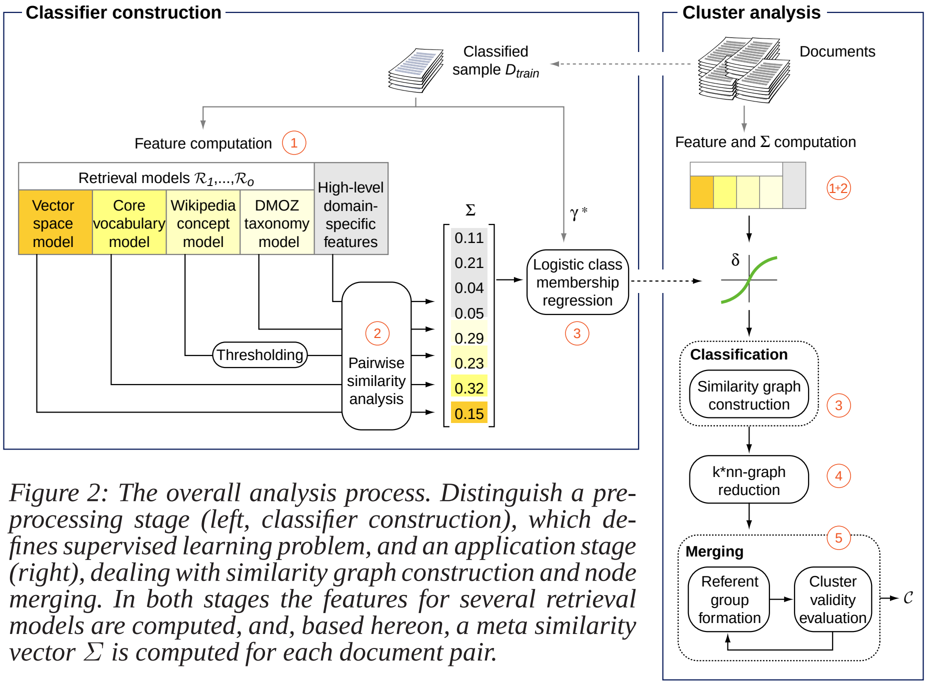

Figure 2: The overall analysis process. Distinguish a pre-processing stage (left, classifier construction), which defines supervised learning problem, and an application stage (right), dealing with similarity graph construction and node merging. In both stages the features for several retrieval models are computed, and, based hereon, a meta similarity vector Σ is computed for each document pair.

(3) Based on $\mathbf{\Sigma}_{train}$ and γ^*_{train}: learn a document pair classifier $\delta : \mathbf{\Sigma} \to [0; 1]$. As classification technology logistic regression was chosen, which provides advantageous characteristics in connection with dichotomous classification problems.

Cluster Analysis:

(1+2) $\forall d_1 \forall d_2 \in D_{test}$: compute the meta similarity vectors $\Sigma(d_1, d_2)$.

(3) Create similarity graph G_δ from the estimated document similarities $\delta(\Sigma(d_1, d_2))$.

(4) Simplify G_δ by a mutual k-nn reduction [Luxburg 2007].

(5) Generate multiple, alternative clusterings with the density-based MajorClust algorithm and choose the best clustering by optimizing the internal validity measure *expected density* [Stein and Niggemann 1999; Meyer zu Eissen 2007].

Recall that the challenge conditions allowed the use of external knowledge at the pre-processing stage (e.g. from Wikipedia), for instance to construct the document pair classifier δ. Afterwards, within in the classification situation for the test set, no usage of online knowledge was permitted.

2.1 The Retrieval Models

Within in the course of the challenge we investigated various retrieval models with respect to their ability to capture knowledge for person resolution: term-based models like the vector space model, text-structure-analyzing models that quantify the text organization, NLP-related models that identify and assess characteristic phrases about named

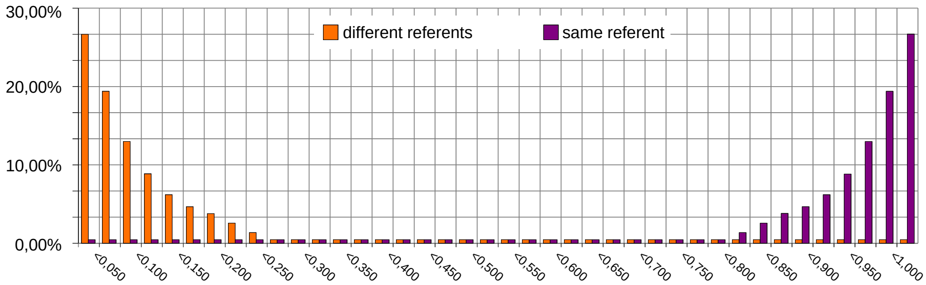

Figure 3: Desired probability distributions for the two classes "different referents" and "same referent" over the range of δ, i.e. the similarity interval $[0;1]$. A combination of ideal retrieval models $\mathcal{R}_1, \ldots, \mathcal{R}_o$ (⇒ Σ^*) and ideal classification technology (⇒ δ^*) will reproduce these distributions: two documents, d_1, d_2, talking about two different referents will get with a high probability a similarity assessment $\delta^*(\Sigma^*(d_1, d_2))$ close to zero. Likewise, two documents talking about the same referent will get with a high probability a similarity assessment close to one.

entities, classification-based models that exploit knowledge compiled within human-edited taxonomies like DMOZ [see 3], and concept-based models such as LSI, pLSI, or ESA, which aim at the identification of hidden connections behind the term surface. After a thorough performance analysis the following models were finally employed:

$\mathcal{R}_1$. Vector space model with $tf \cdot idf$ term weighting and stopwords removed.

$\mathcal{R}_2$. Like $\mathcal{R}_1$, but restricted to the enclosing regions of the target name.

$\mathcal{R}_3$. Match of top-level DMOZ categories using entire document as feature set.

$\mathcal{R}_4$. Like $\mathcal{R}_3$, considering the match of top-level plus second level DMOZ categories.

$\mathcal{R}_5$. Like $\mathcal{R}_4$, but restricted to the enclosing regions of the target name.

This choice does not correspond to the strongest combination to tackle person resolution tasks: in particular, neither NLP-based nor semantically rich models are part of the solution, which is attributed to the runtime constraints that could not be adhered to with the complex retrieval models. The models $\mathcal{R}_2$ and $\mathcal{R}_5$ heuristically accomplish a little of the power of a named entity analysis since they introduce extra weight on the terms in the neighborhood of target names. For the retrieval models $\mathcal{R}_3, \mathcal{R}_4$, and $\mathcal{R}_5$ sophisticated multi-class, multi-label classifiers were constructed.

2.2 Selected Analysis Results and Discussion

Figure 4 shows the probability distributions generated by the document pair classifier δ when the models $\mathcal{R}_1, \ldots, \mathcal{R}_5$ are combined according to an optimum logistic regression of γ^*. The plot may look convincing, but it does not reflect the class imbalance of 25:1 between the document pairs talking about different referents and the document pairs talking about the same referent. When considering this a-priori probability, the averaged precision above a similarity threshold of 0.725 that is generated by δ is approximately 0.2. Stated another way: only every fifth edge in the similarity graph G_δ

[3] DMOZ is a human-edited Web directory, found under `http://www.dmoz.org/`.

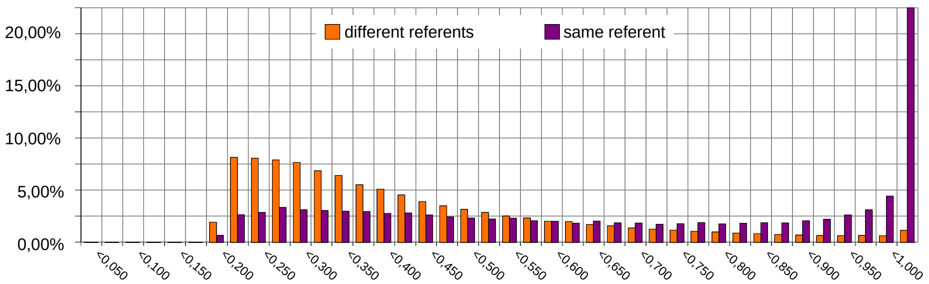

Figure 4: Probability distributions generated by the learned document pair classifier δ, based on the optimized combination of the retrieval models $\mathcal{R}_1, \ldots, \mathcal{R}_5$.

gets a proper weight that reflects the true similarity between its incident documents. Interestingly, this is sufficient to achieve an F-Measure value of 0.42 after the cluster analysis (see Table 1, right). The rationale for this effect is as follows. Edges between documents of different referents are distributed uniformly over G_δ, while edges between documents of the same referent lump together. This characteristic is rooted in the quadratic connection between referent instances (the nodes in G_δ) and the possible relationships between these instances (the edges in G_δ). For a node in G_δ one can compute a—what we call—*edge tie factor*, which quantifies this characteristic as the ratio between correct edges on the one hand, and wrong edges that lead to the *same wrong referent* on the other. This factor, which is between 4 and 6 here, reduces the class imbalance accordingly. Altogether, the class imbalance, the edge tie factor, and the achieved probability distributions for δ (Figure 4) result in 0.65 as the *effective precision* generated by δ above the similarity threshold 0.725.

To capture as much as possible from the δ similarity assessments, the graph G_δ was abstracted by a mutual k-nn reduction, and, state-of-the-art cluster analysis technology was applied, including density-based merging, density analysis, and differential probing. Table 1 compiles the recognition results for the Spock Data Mining Challenge that we finally achieved: the left-hand side shows the F-Measure values for the exclusive use of retrieval models, the right-hand side shows the F-Measure values when using $\mathcal{R}_1, \ldots, \mathcal{R}_5$ in a combined fashion.

Retrieval model (learning: logistic regression)	F_α-Measure $(\alpha = 1/3)$
$\mathcal{R}_1$. *tf·idf*	0.39
$\mathcal{R}_2$. *tf* region	0.32
$\mathcal{R}_3 + \mathcal{R}_4 + \mathcal{R}_5$. DMOZ all	0.15
$\mathcal{R}_6$. ESA Wikipedia persons	0.30
$\mathcal{R}_7$. phrase structure analysis	0.17

Learning technology $(\mathcal{R}_1 + \ldots + \mathcal{R}_5)$	F_α-Measure $(\alpha = 1/3)$
logistic regression	0.42
ensemble cluster analysis	0.40

Table 1: The achieved recognition results for the Spock Data Mining Challenge, depending on the retrieval models and the employed learning technology. Due to performance reasons, $\mathcal{R}_6$ and $\mathcal{R}_7$ were excluded from the final solution.

In terms of the F-Measure the difference between a retrieval model combination and their exclusive use appears small. The main benefit of applying multiple models lies in the improved generalization capability, which could be verified on smaller test sets. Likewise, though the ensemble cluster analysis behaves inferior compared to the logistic regression (see Table 1, right), it has the potential to generalize better when using knowledge-based combination rules.

3 Room for Improvements

From our point of view the most interesting results of the contribution relate to the combination of different retrieval models, and the concept of effective precision. With respect to the former we can clearly state that the full potential has not been tapped; the following promising retrieval models are not part of our solution:

- explicit semantic analysis, ESA [Gabrilovich and Markovitch 2007]
- genre-enabling models [Stein and Meyer zu Eißen 2008]
- named-entity and shallow-parsing models based on a phrase structure analysis

Further room for improvement relates to the field of graph abstraction, where the use of flexible and node-depending thresholds should be investigated.

References

[Bekkerman and McCallum 2005] Bekkermann, R., and McCallum, A.: Disambiguating Web appearances of people in a social network. In *WWW '05: Proc. 14th international conference on World Wide Web*, pp. 463-470, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9.

[Bhattacharya and Getoor] Bhattacharya, I., and Getoor, L.: A Latent Dirichlet Model for Unsupervised Entity Resolution. In Joydeep Ghosh, Diane Lambert, David B. Skillicorn, and Jaideep Srivastava, editors, *SDM*, 2006. ISBN 0-89871-611-X.

[Daumé III and Marcu 2005] Daumé III, H., and Marcu, D.: A Bayesian model for supervised clustering with the Dirichlet process prior. *Journal of Machine Learning Research*, 6:1551-1577, September 2005. URL `http://pub.hal3.name/#daume05dpscm`.

[Finley and Joachims 2005] Finley, F., and Joachims, T.: Supervised clustering with support vector machines. In *ICML '05: Proc. 22nd International Conference on Machine Learning*, pp. 217-224, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: http://doi.acm.org/10.1145/1102351.1102379.

[Fleischman and Hovy 2004] Fleischman, M. B., and Hovy, E.: Multi-Document Person Name Resolution. In *Proc. ACL '04 Workshop on Reference Resolution and its Applications*, 2004.

[Gabrilovich and Markovitch 2007] Gabrilovich, E., and Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. 20th International Joint Conference for Artificial Intelligence*, Hyderabad, India, 2007.

[Luxburg 2007] Luxburg, U.: A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395-416, 2007. ISSN 0960-3174. doi: http://dx.doi.org/10.1007/s11222-007-9033-z.

[Mann and Yarowsky 2003] Mann, G. S., and Yarowsky, D.: Unsupervised personal name disambiguation. In *Proc. seventh conference on Natural language learning at HLT-NAACL 2003*, pp. 33-40, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Meyer zu Eissen 2007] Meyer zu Eissen, S.: *On Information Need and Categorizing Search*. Dissertation, University of Paderborn, Feb. 2007.

[Pasula et al. 2002] Pasula, H., Marthi, B., Milch, B., Russell, S. J., and Shpitser, I.: Identity Uncertainty and Citation Matching. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1401-1408, 2002.

[Stein and Meyer zu Eißen 2008] Stein, B., and Meyer zu Eißen, S.: Retrieval Models for Genre Classification. *Scandinavian Journal of Information Systems*, to appear, July 2008.

[Stein and Niggemann 1999] Stein, B., and Niggemann, O.: On the Nature of Structure and its Identification. *Graph-Theoretic Concepts in Computer Science*, volume 1665 LNCS of *Lecture Notes in Computer Science*, pp. 122-134. Springer, June 1999. ISBN 3-540-66731-8.