

Document Space Workshop Report

Edel Garcia
Mi Islita.com
admin@miislita.com

IPAM, UCLA, California
January 23 – 27, 2006

From January 23 through 27 the applied mathematics community conducted the first ever workshop on document space, at the Institute of Pure and Applied Mathematics (IPAM) in UCLA. Funded by NSF and multidisciplinary in nature, the workshop brought together world experts in Mathematics, Information Retrieval, Statistics, Electrical Engineering, Computer Science and Linguistics.

Thanks to a generous NSF grant from IPAM and to an angel, I was able to attend such unique event. Organized by Carey Priebe, Damianos Karakos, Mauro Maggioni and David Marchette and under the guidance of IPAM's director, Mark Green, the workshop exceeded all expectations.

The workshop was unique in the sense that for the first time the applied math community was able to make their case on what is or is not a document space, a document and a query. It was clear from the start that the days of merely embedding documents using crude vector space, LSI, or mere distance measures are over.

For instance, orthogonal term vector models fail to recognize term-term dependencies, which are essential with semantic associations. While this is addressed to some degree with LSI models, their implementations are not practical with collections consisting of billion of documents. The needs for modeling large and complex datasets in higher dimensions ask for better solutions.

Diffusion geometries and the diffusion space emerged as a good alternative. Not only word-document relationships, but also scalability issues, dynamics of the datasets, and their semantics can be explored and better understood with diffusion geometries.

The Presentations

Day one

IPAM Director, Professor Mark Green, opens day one with a lovely welcome and mentions the mission of IPAM. Carey Priebe (Johns Hopkins) follows. A friendly giant (intellectually and physically), his strong voice reverberates across the audience and makes all feel at home. Carey has an incredible memory –he even remembers my name during the entire event.

Carey introduces the illustrious Michael Trosset from College of William and Mary. His talk, “Trading Spaces: Measuring of Document Proximity and Methods for Embedding Them”, makes clear that similarity and dissimilarity in terms of Euclidean space and Euclidean distances is not enough to represent documents and queries. The general agreement appears to be that queries are linear combinations of terms. However, to define documents one needs to define first the document space.

Professor Trosset reviews binary and quantitative Vector Space Models (VSM), Latent Semantics Indexing (LSI) and Principal Component Analysis. He mentions the difference between the last two as follows: LSI finds the best linear subspace, while PCA finds the best *affine* linear subspace. To find the best affine linear subspace, first translate the feature vectors (y_i) so that their centroid lies at the origin, then find the best linear subspace.

I asked why we need to do this transformation. He mentions that this is done to convert cosine similarities to Pearson’s product-moment correlation coefficients. He then explains the following algorithms: Linear Discriminant Analysis, Guttman Majorization Algorithm and Diagonal Majorization Algorithm.

We stop and I go to lunch with a group consisting of Michael Berry, David Marquette, John Conroy and others I don’t remember their names. Sorry.

The afternoon session starts with Michael Berry (Univ. of Tennessee), co-author with Murray Browne of the upcoming book *Lecture Notes on Data Mining* (World Scientific). Michael, along with Susan Dumais and Tom Landauer, is considered a pioneer of LSI.

His talk, “Text Mining Approaches for Email Surveillance”, demonstrates the beauty of NMF and LSI when applied to email forensics. His test case was the now famous Enron Corpus, an email set consisting of more than 500,000 email messages generated between 1999 and 2001. This was a corpus edited by William Cohen (CMU) from an original set of 15 million email messages.

Berry's procedure is elegant:

1. parse inbox and private folder of all 150 accounts (users) via GTP (General Text Parser).
2. use a 495-term stoplist
3. extract terms appearing in more than 1 email and more than once globally.
4. Then construct a term-message matrix to assign term weights using an entropy model, Non Negative Matrix Factorization (NMF) and Gradient Descent-Constrained Least Squares (GDCLS).

This approach grouped specific topics to dominant terms. For instance the topic "Enron collapse" grouped the following terms: Enron partnership(s), collapse, fastow, shares, sec, stock, shareholder, investors, equity, lay. Amazing research.

Sanjeev Khundapur (Johns Hopkins) is next with "Document Representations for Topic-Adaptation in Statistical language Modeling". Sanjeev presents several maximum-entropy models that seem to be a reasonable way to incorporate topic-information into a statistical language model for, among others, speech recognition.

These models treat word sequences as realizations of a Markov (finite memory) process. A word-document co-occurrence matrix is used in which weighted word co-occurrences are computed. In my opinion this presentation was a diversion from the main topic of the workshop; i.e., document space models and applications. Nevertheless, it was great music to my ears and I didn't mind since I'm much into word co-occurrence theory.

The last speaker of the day was the legendary Ronald Coifman (Yale, Diffusion Geometries Group). In his presentation, "Diffusion geometries of digital document spaces, ontologies and knowledge building", he explained why with multiple dimensions and large collections the notion of distance is meaningless.

There is a better way for embedding documents and modeling them after all: diffusion geometries and mapping in the diffusion space. He explains how they use self-similar, multidimensional scaling and a simple technique of radio 1, radio 2 in the analysis. Applications to collaborative filtering are also discussed.

Day Two

The workshop continues with a back-to-back presentation on Topic Models by John Lafferty's (Carnegie Mellon) and David Blei (Princeton University), inventor of Latent Dirichlet Allocation (LDA).

"Topic Models 1: Probabilistic Models of Documents and Topic Models 2: Structured and Dynamic Models" presentation starts. In part one they cover how low dimensional representations of doc collections can be applied to Collaborative Filtering, IR, Anotate unlabeled images and for topic evolution over time.

Latent Dirichlet Allocation (LDA) is discussed. Blei explains that each document is considered a random mixture of topics from which documents can be generated. LDA is used with "bag of words" for several reasons. One is to identify the themes of documents. The other is to find the most representative words from specific topics. Useful applications to image retrieval are also presented.

Next is Eugene Charniak (Brown University), with "Recent Results form Parsing". Prof Charniak is a venerable parser expert. He explains that parsing is the problem of mapping a string to a phrase structure. He gets into precision and recall issues from the parsing standpoint; then defines what is a constituent and constituent accuracy.

Charniak also reviews the history of parsers up to current days and then gets into parsers cross validation and self-training results. Personally, I found his presentation quite off-topic from the central topic of the workshop.

Frederick Jelinek (Johns Hopkins), a brilliant man, is next with "Experiments with Random Forests". He explains why a language model is a distribution over words and gets into a model based on words occurrence and acoustic.

Jelinek explained that some use the frequency of a word divided by the frequency of a previous word as a smoothing ratio and why this is a poor estimate of probability. He mentions that Kneser-Ney Smoothing, an iterative technique for smoothing data should be used. He then present examples of random decision trees and some drawback of these trees.

Jason Eisner (John Hopkins Univ) delivers perhaps the most electrifying presentation: "Bootstrapping without the Boot". He discusses how to tackle the double spiral problem when we do clustering. Then presents an unsupervised learning technique that requires little math. It is a clever technique to address the problem of unsupervised WSD (word sense disambiguation). It consists in mining results using pseudo words as initial seeds from which a lexical tree is

constructed. These are words created by connecting any two terms. Clever idea. I can see many applications with On-Topic Analysis.

Day Three

Damianos Karakos (John Hopkins Univ) presents "Language Model with the Maximum Likelihood Set: Complexity Issues and the Back-Off Formula".

He mentions that the Maximum Likelihood Set (MLS) was introduced as a parameter-free technique for estimating a probability mass function (pmf) from sparse data. The MLS contains all pmfs that assign merely a higher likelihood to the observed counts than to any other set of counts, for the same sample size.

He explains that language modeling is a probabilistic assignment to any word sequence and then mentions applications in the area of document categorization, information retrieval and speech recognition.

Peter Jones (Yale Univ) is next with "Eigenfunction Local Coordinates and the Local Riemann Mapping Theorem." He explains that a geodesic distance is the shortest path between two diffusion points. As mentioned in his abstract and quote:

"One idea in the exciting new area of Diffusion Geometry is to use certain eigenfunctions as new local coordinates on a data set (e.g. a collection of documents). These coordinates are surprisingly robust under perturbation of the underlying sets and have been empirically observed to provide local coordinates on rather large patches."

He then explains that this robustness is a "hidden" feature of the Riemann Mapping Theorem for simply connected planar domains that is quite general. Robustness also works on manifolds of arbitrary dimension, Jones stated. His findings reinforce the notion that robustness in complex systems is a self-similar, scale-independent feature.

The next speaker is David Marquette from the Naval Surface Warfare Center. His presentation, "How Document Space is like an Elephant?" is a work in collaboration with NAVSEA and Johns Hopkins University.

David explains the many views of scientists (applied mathematicians, statisticians, IR, et) when it comes to defining what is a document space, from here the title of his presentation. He mentions that documents should be represented for tasks like classification, clustering, summarization, etc.

When mapping document features into a space he recommends:

1. not to extract more features than those needed,
2. use the simplest method.
3. do iterations (feedback).

He then discusses the so-called Dimensionality Reduction Curse of LSI and SVD (singular value decomposition).

The next speaker is Michael W. Mahoney from Yahoo! with "Data-driven Dictionary Definition for Diverse Document Domains." Mahoney reviews LSI's SVD and introduces new advances in SVD and matrix selection, mentioning that term-document data, recommendation system data, individual-gene data, and temporal image data can be represented as term-document matrices.

Essentially his work focuses on applying SVD and data matrix analysis to dictionary mapping. He describes a low-rank matrix decomposition that is expressed in terms of a small number of actual rows and actual columns.

The next speaker of the day is Andrew Tomkins, also from Yahoo! with "Representation of Web Document Spaces".

Andrew revisits the infamous BowTie Theory Graph of the Web, a work he did with Andrei Broder back in 2000. I asked a question about the current state of this graph and he mentions that the jury is still undecided but that overall the OUT link part of the graph is growing bigger than the other two (IN and SCC).

He then discusses how and why is important to identify specific link structures and gives examples using a K2,3 Web community. This is a specific connection between nodes consisting of two hubs and three authorities, like two pages both pointing to any and same three pages. A core $K_{i,j}$ then consists of i left nodes, j right nodes and all left-right edges, he explains.

Why studying connectivity between cores is important? Well, he argues that the shortest path alone is not always a good choice. At this point he gets into an electrical and atomic model to illustrate the problem of flow and "voltage" between nodes. One problem in web subgraph analysis, he mentions, is to find a subgraph that captures much of the flow. His diagrams remind me of voltage relaxation problems found in potential theory books.

Day Four

The first two are back-to-back presentations from Coifman's Diffusion Geometries Group. The first speaker is Stephane Lafon now at Google and Mauro Maggioni (Yale). Lafon's topic is "Geometric clustering in kernel embedding spaces for document corpora organization". This is a collaborative work between Google, Ronald Coifman's Diffusion Geometries Group at Yale, Princeton, Carnegie Mellon and Weizmann.

In my opinion, diffusion geometries is the future of document space embeddings. Stephane explains that collections of documents are associated to a graph whose structure is analyzed and organized via "diffusion coordinates". As his abstract states and quote:

"In the corresponding embedding space, one can perform simple geometric algorithms, such as k-means clustering or ball covering, in order to obtain a meaningful organization of the corpus. The dual approach is to consider the set of words contained in these documents as the data of interest, leading to an automatic lexicon analysis and concept extraction scheme."

"The problem of finding an "optimal" clustering of the documents in the embedding space is reduced to simple questions of matrix approximation and completion. This provides a rigorous justification for kernel k-means. In addition, this idea goes beyond the diffusion framework and classical symmetric kernel setting, as it allows to deal with arbitrary oriented graphs and kernels."

Since data sets exist in higher dimensions, dimensionality reduction techniques are necessary. This is a challenge. Another challenge, he mentions, is that we need to presort results (pre-rank) or do clustering in the diffusion space. A third challenge of this work is how to extend the model to non-symmetric spaces.

Collections of documents are described in terms of a diffusion graph with diffusion coordinates. The structure of this graph is analyzed using k-means clustering, covering balls (radio 1, radio 2 balls) or other geometric algorithms. I'm of the opinion that this is where scaling techniques (fractals) helps to address the complexity of the diffusion graph.

Stephane then describes a circuital and molecular model Google is using to model document categorization. It would be interesting to know if they have thought about developing rotational selection rules for their "atoms" and "molecules". In my opinion, Lafon and Coifman presentations were the best presentations of the workshop.

Next is Mauro Maggioni. As the previous one, his topic, "Multi-scale Analysis of Graphs and Document Corpora" is equally fascinating. His abstract summarizes the presentation very well and quote:

"Diffusion processes and random walks on graphs allow to define and construct multi-scale structures on graphs, for example arising from high-dimensional point clouds, or bodies of documents. This coherent multi-scale organization allows one to analyse the graph at different levels of resolution, to reveal (soft) clusters and communities, and to construct multi-scale learning algorithms."

"When the graph is associated with a body of documents, this construction leads to two dual, tightly related, multi-scale structures, one on documents and one on words and concepts, which allow to extract information at different levels of specificity."

No doubt this was a great topic. However, from the presentation standpoint, it was not of good quality. The visual slides were simply copies in article format, hard to read and to follow while the speaker talked. The audience kept staring at paragraphs after paragraphs with few captions and visual aides, except for the figures embedded in the article. It could has been delivered in a better format.

The next speaker is David Horn (Tel Aviv University). His presentation, "Unsupervised learning of Natural Languages" tries to address the following: a document is a collection of symbols from which the underlying rules that govern its production can be inferred. For this purpose he uses ADIOS (Automatic Distillation of Structure), their pattern extraction algorithm. In his presentation abstract, he claims and quote:

"This is the first time an unsupervised algorithm is shown capable of learning complex syntax, generating grammatical novel sentences, and proving useful in other fields that call for structure discovery from raw data, such as bioinformatics."

In my opinion, his presentation was relevant to linguistics and bioinformatics, but can hardly be said that it was relevant to the main topic of the workshop; i.e. the Document Space.

The next speaker in line is Nello Cristianini from UC Davis. He presents on "Kernel Methods for Text Analysis" An expert in kernel methods, Nello shows how these can be used with many algorithm, including document space embedding models. He provides a long list of applicable algorithms, explaining how kernel methods can be used to embed text from documents in a semantically meaningful way. Applications to PCA, LSI, CCA and string matching algorithms are also provided.

Next is Piotr Indyk from MIT. He presents on "Algorithmic Applications of Low-Distortion Embeddings". Essentially he present a review on the history of embedding maps and models. In particular he covers dimensionality reduction and near neighbor metrics for computer vision.

Day Five

The last day of the workshop and only two presentations were scheduled.

John Conroy (IDA Center of Computing Sciences) starts with "Multi-Document Summary Space:What do People Agree is Important?". John states that a multi-document summary gives the "gist" of what is contained in a collection of related documents, i.e. what a collection is about.

To address the question of how to define a "gist" he estimates the probability that a word selected by a human will be included in written summaries of document sets.

But how can we define a "gist?" His abstract states:

"We explore this question by analyzing human written summaries for clusters of document sets. In particular, we estimate the probability that word will be chosen by a human to be included in a summary. We demonstrate that if this probability model were given by an oracle, then a simple automatic method of summarization can extract summaries which are statistically indistinguishable from the human summaries."

To understand this, imagine that you conduct a search in Google and gathers the top N documents. Then you write a summary describing the content of those top N documents. This summary contains the gist of the N documents relevant to the input query. He argues that a human editor or an automatic method for extracting such summaries can produce summaries that are statistically indistinguishable.

The last speaker of the workshop is Djoerd Hiemstra from Universiteit Twente. His presentation "Expressing language modeling approaches as region algebra queries" proposes a unified theory of document space by combining two different approaches:

1. region models = developed for structured document retrieval.
2. language models = useful for ranking search results and for developing new ranking approaches.

Region models provide a structured query language while language models are used to justify ranking of search results, and for developing new ranking methods.

Hiemstra claims that there is a one-to-one relationship between region queries and the language models they represent for a wide variety of applications such as simple ad-hoc search, cross-language retrieval, video retrieval, and web search.

Workshop ends.