Populating Categories using Constrained Matrix Factorization

Read the Web Project Report

Andreas Zollmann

January 25, 2010

Abstract

Matrix factorization methods are a well-scalable means of discovering generalizable information in noisy training data with many examples and many features. We propose a method to populate a given ontology of categories and seed examples using matrix factorization with constraints, based on a large corpus of noun-phrase/context co-occurrence statistics. While our method performs reasonably well on some categories, it is outperformed by a simple nearest-neighbor based baseline. We demonstrate, however, that dimensionality reduction applied to the baseline model improves performance considerably.

1 Introduction

The task of macro-reading is to learn facts about the world, such as properties of and relations between entities, from a large corpus of text. In this work, we restrict ourselves to the scenario of populating an ontology of predefined categories with noun phrases, and assume that the text corpus can be summarized by a matrix in which rows index potentially relevant noun phrases, columns index potential context phrases, and each entry expresses the number of times the respective noun phrases occurred in the respective context. Given a small number of 'seed' example noun phrases for each category, our task is to find a given number of noun phrases for each category of the ontology with as high precision as possible.

Matrix factorization methods are a well-scalable means of discovering generalizable information in noisy training data with many examples and many features. We propose to populate the ontology using matrix factorization with constraints, similar in spirit to singular value decomposition

(SVD). Our training data, a noun-phrase-to-context co-occurrence matrix of dimension $100,000 \times 100,000$, is large but sparse; therefore it is crucial that our algorithms work on sparse matrices directly.

2 Related Work

Large scale matrix factorization is a well studied problem (especially since the introduction of the Netflix problem to the learning community).

There is a vast amount of literature on matrix factorization with constraints, cf. e.g. (Yang and Hu, 2007; Hoyer, 2004; Zheng et al., 2007). While many have studied incorporating sparsity constraints (representing each data point in the original space with *few* basis vectors), especially for the case of non-negative matrix factorization (NMF), and although we were motivated by their approaches to incorporating constraints in general, we wish to exploit different kinds of constraints (ones defined by our ontology), and a different kind of sparsity (sparsity of the *original* data matrix.

Four different methods of singular value decompositions based on sparse data matrices are described by Berry (1992). Dealing with data matrix sparsity was crucial during the Netflix competition of predicting users' movie ratings as well, where Funk (2006) was the first to come up with an efficient matrix factorization algorithm, variants of which were subsequently used by all the leading teams (Takács et al., 2009). In essence, this algorithm minimizes the mean squared error of all nonzero entries in the sparse user-by-movie rating matrix by performing stochastic gradient descent based on these nonzero entries. As the zero entries in the rating matrix represent unknown values, this is the correct objective. In our scenario, zero-entries actually represent known values, so this technique is not directly applicable. We nevertheless choose a gradient-descent based approach, which makes it easy for us to relax the orthogonality constraint and incorporate our ontology-specific constraints.

We also found Yu et al. (2009) to be both interesting and relevant. The paper introduces an EM style algorithm to solve a problem similar to the one we formulate and incorporates several tricks to compute the desired E and M step updates efficiently.

3 Methods

3.1 Formulation

Given co-occurrence statistics of noun phrases and their contexts in a web corpus, we want to project the noun phrases, viewed as training examples whose features are the possible contexts, down to a lower-dimensional space, by optimizing the objective

$$\min \|A - \hat{A}\|_{\mathcal{F}} \tag{1}$$

where A is the given NP-context matrix, $\hat{A} = U\Sigma V^T$ is a low rank approximation to A, and $\|.\|_{\mathcal{F}}$ denotes the Frobenius norm. We can "absorb" the singular values in Σ into the matrix V and write $\hat{A} = UV^T$.

Let us look closer at the SVD, the matrix U is simply the projection of the input NPs into an orthogonal basis (possibly of lower dimension). Given a set of seed vectors $\mathcal{S}$, with known categories, we wish to additionally constrain our matrix factorization so that for each seed vector s, its projection $(x_1, \ldots, x_N)$ (a row in U) must be of the form (where $\{c_k | k \in cat(s)\}$ are the categories of the seed vector s):

$$x_k > 0, \text{ for all } k \in \text{ cat}(s) \tag{2}$$

$$x_i = 0, \text{ for all } 1 \leq i \leq K \text{ with } i \notin \text{ cat}(s) \tag{3}$$

We can then assign a new non-seed NP with projection $(x_1, \ldots, x_N)$ to categories by choosing all c_k for which x_k exceeds a given threshold. Unlike in SVD, the solution to this problem is not necessarily orthogonal (i.e. because of the constraints we do not expect the columns of U or V to be orthogonal). We instead add a "non-orthogonality penalty" to bias our solutions towards orthogonality explicitly (it is desirable in general for the basis vectors to be as "far" from each other as possible, in terms of cosine distance, so that our classification into categories is robust).

To do this we reformulate our objective as

$$\min \|A - UV^T\|_{\mathcal{F}}^2 + \alpha \sum_i \sum_{j \neq i} (V_i^T V_j)^2 \tag{4}$$

where m is the number of noun-phrases and n is the number of contexts in our original matrix, and α is a parameter controlling the tradeoff between the reconstruction error and orthogonality of the basis. V_i and V_j refer to the i^{th} and j^{th} columns of the matrix V. We also add to the objective function the squared Frobenius norms of U and V as a regularization penalty

parameterized by β, primarily to ensure numerical stability and to avoid circumvention of the non-orthogonality penalty by decreasing the norm of columns in V and increasing the norm of the respective columns in U accordingly. In this case we can simplify the non-orthogonality penalty by including the diagonal products $V_i^T V_i$, yielding the final objective:

$$\min \|A - UV^T\|_{\mathcal{F}}^2 + \alpha(V^T V)^2 + \beta(\|U\|_{\mathcal{F}}^2 + \|V\|_{\mathcal{F}}^2) \tag{5}$$

Note that V is now regularized more strongly than U than it would be with the original non-orthogonality penalty.

3.2 Solving the optimization problem

Perhaps the most obvious barrier to solving this problem directly is that even the smaller (roughly 100Kx100K) noun-phrase/context co-occurrence count matrix M would require 40GB of memory when stored directly in memory (assuming 4 bytes per entry). Therefore working with sparse matrices is crucial.

We implemented an iterative gradient descent algorithm (optimize over entries in U, and then in V iteratively), to solve the problem. We note however that the matrices U and V are significantly smaller (but dense). The gradients of V for our objective above are giving by

$$2A^T U - 2VU^T U + 4\alpha V V_T V + 2\beta V \ .$$

The U-gradients are analogous, except that they do not have the α term. In contrast to the objective itself, whose computation is of time complexity $O(MN)$, where M is the number of noun phrases and N the number of contexts, the gradients can be computed in $O(M + N)$ time, if the number of nonzero elements in a row or column of A is bounded by a constant.

4 Experiments

For our experiments, we selected ten categories from the Read the Web ontology, together with their seed noun phrases, to build the model. Apart from the ten dimensions corresponding to the categories, we used ten additional unconstrained dimensions. We preprocessed the noun-phrase by context matrix with the sparsity-preserving element-wise $\ln(1 + x)$ operation, which gave better results than using raw counts.

We compare our model against two nearest-neighbor based baselines: *Baseline-direct* suggests for each category the noun phrases that are closest

to the mean of the seed noun phrases of that category. *Baseline-SVD100* uses the same selection criterion, but first reduces the feature space from the 100K contexts to a 100-dimensional space using singular value decomposition. We tested cosine distance as well as Euclidean distance, the latter of which worked much better. Again, we applied the $\ln(1+x)$ preprocessing step, which led to considerably better results than raw counts.

The following results show, for each category, the twenty best noun phrases, excluding the seed noun phrases, suggested by the three methods.

- Academic field
  Baseline-direct: Anthropology, Political Science, Biochemistry, Art History, Classics Mechanical Engineering, International Relations, Geography, Electrical Engineering, Criminal Justice, Civil Engineering, Religious Studies Chemical Engineering, comparative literature, studio art, English Literature, Journalism, cultural anthropology, Botany, zoology
  Baseline-SVD100: Chemistry, Anthropology, chemical engineering, Geography, civil engineering, mechanical engineering, Engineering, clinical psychology Political Science, business administration, Accounting, English literature, pharmacology, electrical engineering, zoology, biomedical engineering, Finance, microbiology, environmental science, astrophysics
  Constrained MF: mathematics, psychology, economics, physics, biology, history, chemistry, computer science, English, philosophy, science, education, engineering, music, business, law, sociology, medicine, anthropology, art

- Sport
  Baseline-direct: rugby, bowling, softball, snowboarding, lacrosse, paintball, table tennis, ice hockey, skating, rock climbing, croquet, kayaking, rodeo gymnastics, archery, skateboarding, rowing, college basketball, drag racing, cheerleading
  Baseline-SVD100: cricket, chess, surfing, opera, fly fishing, country music, NASCAR hip-hop, boating, Jazz, mountain biking, birding, biking, anime, poker ballet, gymnastics, songwriting, baking, kayaking
  Constrained MF: life, music, the game, sports, work, business, research, politics, golf, training, teaching, the world, education, art, fishing, writing, food, the sport, people, science

- Politician
  Baseline-direct: Al Sharpton, George W Bush, Yushchenko, Condi, Keyes, Nancy Pelosi Gravel, GWB, Santorum, Sanjaya, Brownback, Senator McCain, JM, Pat Buchanan, Chirac, Bill Richardson, Lamont, Peres, Newt Gingrich Michelle Obama
  Baseline-SVD100: Mitt Romney, Barack, Mike Huckabee, Senator Obama, Nader, John Edwards, Giuliani, Rove, Lieberman, Senator Clinton, Putin, Steve Jobs, Fred Thompson, Howard Dean, Karl Rove, Rumsfeld, John Kerry Tony Blair, Gordon Brown, Castro
  Constrained MF: Obama, Bush, McCain, Clinton, Hillary, God, Jesus, President Bush, Paul, people, Kerry, John, Mr, the man, Romney, the president, John McCain, the President, the person, a man

- Actor
  Baseline-direct: Qaida, Hast, Rocky Mountains, Pero, vBSEO, das, This seller,

Smaller voor, minded individuals, Highest, homelite chain, university faculty members, Dynamically, Method Detail, Recent years, gmail dot com ligne, Improperly, little tikes
Baseline-SVD100: Kate Moss, Cameron Diaz, Mick Jagger, Jennifer Aniston, Charlton Heston, Dolly Parton, Fabio, Bogart, Shakira, John Travolta Tarantino, Jane Fonda, Ben Affleck, Jim Carrey, Russell Crowe Stallone, Steve Martin, Pamela Anderson, Alex Rodriguez, Don Imus
Constrained MF: actor, actress, a woman, a man, Mr, Brad Pitt, John, the actor, Tom Cruise, the man, Johnny Depp, Dr, the guy, David, people, a girl, Angelina Jolie, Chris, Peter, Michael

- Athlete
  Baseline-direct: Qaida, Hast, Pero, das, Rocky Mountains, vBSEO, This seller, voor minded individuals, Smaller, Lowest, ligne, Highest, homelite chain university faculty members, Recent years, Dynamically, gmail dot com Method Detail, little tikes
  Baseline-SVD100: Ziggy, Sosa, Shakira, Snoop, Don Imus, Marlena, Bruce Willis, Joe Torre, Arnie, Gabby, Em, Fidel, Ally, Mohinder, Mika, Ben Affleck Bill Murray, Stefano, Boo, Tendulkar
  Constrained MF: father, Jesus, mother, daughter, son, husband, dad, God, the man, brother, Jack, David, John, The house, sister, a man, Michael, Peter, Bush, wife

- Movie
  Baseline-direct: Saving Private Ryan, vBSEO, Tornado, The Ring, This seller, Pero Rocky Mountains, Hast, das, little tikes, Qaida, Toy Story, voor minded individuals, Smaller, gmail dot com, ligne, Highest, Recent years, Method Detail
  Baseline-SVD100: Lord of the Rings, Final Fantasy, Transformers, Spiderman, Jurassic Park, Doom, Battlestar Galactica, the original film, Zelda, Guitar Hero, the Lord of the Rings, Cloverfield, Oblivion, Monopoly, The Sims, Romeo and Juliet, X-Men, King Kong, Firefly, Sailor Moon
  Constrained MF: the movie, the film, the book, the show, the story, the game, the series, the play, this film, a movie, this book, this movie, a book, the video, the song, the world, life, the work, today, books

- Animal
  Baseline-direct: lizards, salamanders, iguanas, dragonflies, sparrows, earthworms tortoises, hawks, herons, crayfish, hippos, starfish, toads, terns lobsters, scorpions, pheasants, kangaroos, caterpillars, sea birds
  Baseline-SVD100: turtles, frogs, spiders, fishes, penguins, bats, bears, lions elephants, orchids, wild animals, goats, reptiles, flies, chickens seals, parrots, monkeys, mosquitoes, trout
  Constrained MF: people, animals, fish, children, plants, students, men, women, things, trees, insects, dogs, individuals, books, items, products, objects, horses, food, water

- Company
  Baseline-direct: Hewlett-Packard, Hewlett Packard, Nortel, Vodafone, Ericsson Qualcomm, Xerox, Siemens, Texas Instruments, Cisco Systems, Raytheon RealNetworks, Lockheed Martin, SGI, Autodesk, Symantec, Nestle Lucent, Philips,

Westinghouse
Baseline-SVD100: Dell, Sony, Nintendo, Nokia, HP, Toyota, Cisco, Adobe, AOL, ABC Verizon, CBS, Motorola, Wal-Mart, NBC, Sun, Novell, Nike, Honda, SAP
Constrained MF: the company, Google, people, the government, God, the site, a number, a company, the state, IBM, Yahoo, this site, students, Mr, the University, companies, Dr, the Internet, the group, the people

- City
  Baseline-direct: Minneapolis, San Antonio, Cincinnati, Salt Lake City, Kansas City Munich, Milwaukee, Calgary, Tampa, Sacramento, Birmingham, San Jose Shanghai, Tucson, Ottawa, Albuquerque, Winnipeg, Buffalo, Montreal Madrid
  Baseline-SVD100: Tokyo, Melbourne, Detroit, Beijing, Berlin, Dublin, Montreal Nashville, Baltimore, Amsterdam, Vancouver, Denver, Miami, Athens Sydney, Portland, Cleveland, Dallas, NYC, Vienna
  Constrained MF: New York, Paris, Los Angeles, San Francisco, the city, Washington, New York City, the area, Seattle, order, town, California, Europe, the United States, Atlanta, Toronto, England, St, Philadelphia, Japan

- Country
  Baseline-direct: Argentina, Belgium, Denmark, Finland, Malaysia, Switzerland, Hungary, Norway, Sweden, Nigeria, the Netherlands, Peru, Chile, the Philippines, Romania, Poland, Sri Lanka, Portugal, Kenya, Holland
  Baseline-SVD100: Brazil, Turkey, Thailand, Greece, Poland, Pakistan, Korea, the Philippines, Sweden, Cuba, Scotland, Taiwan, the Netherlands, the South, Malaysia, Singapore, Hong Kong, Norway, the United Kingdom, Kenya
  Constrained MF: the United States, Europe, the world, India, the country, China, Australia, the area, order, Japan, America, England, the city, the UK, California, the region, the U.S., the state, Italy, New York

As quite evident from qualitative inspection, our method, while performing reasonably well on some categories, such as academic fields, is clearly outperformed by both baselines. It further tends to generalize too much, adding high-level concepts, such as "actor" to the actor category, or "the movie" and "the film" to the movies category.

Amongst the baselines, *Baseline-SVD100* performs better. Looking at the actors and movies categories for example, *Baseline-direct* falsely suggests many (presumably rare and therefore badly estimated) foreign-language terms such as "Hast", "das", "voor", "ligne", while *Baseline-SVD100* gets nearly all of them right. We believe that this is due to the sparsity of the training examples in the original noun-phrase/context space, which is ameliorated by the dimensionality reduction.

4.1 Interpretation of the Learned Factors

Tables 1 and 2 show the relatively most influential noun phrases and contexts for different factors of the standard SVD and the constrained matrix factor-

ization algorithm, respectively. We define 'most influential noun phrase i' for factor k as the one for which

$$\frac{|U_{i,k}S_{k,k}|}{||U_{i,.}S||_2}$$

The division of each entry by the respective row norm ensures the exclusion of common noun phrases that have high U-values for all factors. We further exclude all noun phrases appearing less than $10,000$ times from the ranking. We determine the 'most influential contexts' analogously.

We can see that the SVD factors represent the data according to concepts that are often hard to interpret, whereas the constrained MF method is forced to adhere to the predetermined categories for the first 10 factors. For 'academic field', 'sports', and 'city', it manages to find discriminative contexts, and achieves high confidence, as can be seen by the magnitudes, which are close to one. For 'country', on the other hand, the magnitudes are lower. The unconstrained 11-th dimension appears to be mainly have the function of fine-tuning the seed examples, providing negative evidence for certain cities (which are all seeds of the ontology). The other unconstrained dimensions behave similarly, which is a sign that our method overfits to the seed noun phrases.

5 Conclusions and Possible Future Work

We presented a matrix-factorization approach to categorize noun phrases based on the contexts they appear in. Our empirical results are discouraging when compared to those of a simple nearest-neighbor based baseline. We demonstrate, however, that singular value decomposition of the input space can improve that baseline considerably.

Our framework allows for the incorporation of additional "ontological constraints". Consider, for example two constraints of the following form:

- if in category A, must be in category B
- if in category C, cannot be in category D

Now we can easily add these constraints into the original formulation by two additional penalities, of the form:

$$P_1 = \|U_A - U_B\|_{\mathcal{F}} \qquad P_2 = -\|U_C - U_D\|_{\mathcal{F}} \tag{6}$$

where U_A etc. are as defined previously. Further, it is conceivable to extend the model to work with relations by suitably modifying the objective and final classification procedure.

It is worth noting that the SVD-enhanced baseline is very strong indeed: A manual inspection of the 1000 highest-rank suggestions showed impressively high accuracy for most of the categories. This approach could therefore be used as a complementary module suggesting candidates in the Read-the-Web system.

6 Acknowledgements

We thank Sivaraman Balakrishnan for his initial contributions to this project and many useful suggestions.

References

Michael W. Berry. Large scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6:13–49, 1992.

Simon Funk. Netflix update: Try this at home. http://sifter.org/∼simon/journal/20061211.html, 2006.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints, Aug 2004.

Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Scalable collaborative filtering approaches for large recommender systems. *J. Mach. Learn. Res.*, 10:623–656, 2009. ISSN 1533-7928.

yu-Jiu Yang and bao-Gang Hu. Pairwise Constraints-Guided non-negative matrix factorization for document clustering. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 250–256. IEEE Computer Society, 2007.

Kai Yu, Shenghuo Zhu, John Lafferty, and Yihong Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 211–218, New York, NY, USA, 2009. ACM.

W.S. Zheng, S.Z. Li, J.H. Lai, and S.C. Liao. On constrained sparse matrix factorization. In *ICCV07*, pages 1–8, 2007.

Highest relative-magnitude $U \times S$ values	Highest relative-magnitude $V \times S$ values
Factor 1	
all sorts (0.9971)	followed by _ (0.995108)
a range (0.992638)	known as _ (0.994256)
radio (0.99138)	think of _ (0.98952)
plenty (0.991251)	_ consisted of (0.987388)
Java (0.990695)	knowledge of _ (0.987335)
Factor 2	
the midst (-0.749129)	the north of _ (-0.79625)
the basement (-0.744949)	the south of _ (-0.793643)
August (-0.74314)	was killed in _ (-0.768902)
the middle (-0.740454)	the outskirts of _ (-0.7666)
November (-0.740378)	is located in _ (-0.762893)
Factor 3	
Students (-0.792819)	a small group of _ (-0.805571)
Members (-0.772438)	_ are looking for (-0.79151)
GNU FDL (0.761778)	_ benefit from (-0.790194)
scholars (-0.746071)	A group of _ (-0.78171)
People (-0.737509)	_ interested in (-0.779313)
Factor 4	
the Board of Directors (0.694539)	the discretion of _ (0.741127)
the Committee (0.67588)	A friend of _ (0.72413)
the Board (0.670986)	permission of _ (0.715332)
the instructor (0.658759)	'm sure _ (0.669893)
the director (0.643428)	_ graduated from (0.669606)
Factor 5	
top (-0.716642)	the links on _ (-0.749949)
the right side (-0.703272)	contained on _ (-0.727828)
a regular basis (-0.702042)	are located on _ (-0.718561)
the left side (-0.694558)	print out _ (-0.717877)
the verge (-0.69442)	feelings of _ (0.717719)

Table 1: Analysis of first principle components of standard SVD.

Highest relative-magnitude $U \times S$ values	Highest relative-magnitude $V \times S$ values
Factor 'Academic field'	
Arts (0.927776)	methods of _ (0.998768)
computer science (0.877104)	completion of _ (0.99865)
psychology (0.827268)	a professor of _ (0.998342)
chemistry (0.80684)	_ needed for (0.99825)
biology (0.802234)	a degree in _ (0.99818)
Factor 'Sports'	
baseball (0.994758)	a time of _ (0.99995)
football (0.991261)	the act of _ (0.999938)
golf (0.89225)	a lifetime of _ (0.999918)
fishing (0.841256)	rules of _ (0.999842)
the sport (0.83157)	rate of _ (0.999841)
Factor 'City'	
Chicago (0.969319)	the outskirts of _ (0.996735)
Denver (0.96782)	the city of _ (0.996418)
Tokyo (0.966862)	sat in _ (0.99403)
Berlin (0.966317)	the works of _ (0.993579)
Atlanta (0.964662)	located at _ (0.993062)
Factor 'Country'	
South Africa (0.959138)	takes into _ (-0.975073)
New Zealand (0.94867)	_ are provided for (-0.965545)
Indonesia (0.946262)	has served as _ (-0.962099)
Iraq (0.939331)	be put in _ (0.961734)
Mexico (0.927709)	was during _ (-0.960996)
Factor 11	
Moscow (-0.409041)	standing on _ (-0.336474)
Houston (-0.353103)	_ enrolled in (0.321818)
Baghdad (-0.30461)	lying on _ (0.315253)
London (-0.247926)	a spirit of _ (0.308692)
Boston (-0.212334)	decide on _ (0.307499)

Table 2: Analysis of some well- and some badly-learned factors of the constrained matrix factorization method.