Finding Your Way in a Multi-dimensional Semantic Space with Luminoso

Robert Speer
MIT Media Lab
Cambridge, MA, USA
rspeer@mit.edu

Catherine Havasi
MIT Media Lab
Cambridge, MA, USA
havasi@mit.edu

Nichole Treadway
MIT EECS
Cambridge, MA, USA
knt@mit.edu

Henry Lieberman
MIT Media Lab
Cambridge, MA, USA
lieber@media.mit.edu

ABSTRACT

We present Luminoso, a tool that helps researchers to visualize and understand a dimensionality-reduced semantic space by exploring it interactively. It also streamlines the process of creating such a space, by inputting text documents and optionally including common-sense background information. This interface is based on the fundamental operation of "grabbing" a point, which simultaneously allows a user to rotate their view using that data point, view associated text and statistics, and compare it to other data points. This also highlights the point's neighborhood of semantically-associated points, providing clues for reasons as to why the points were classified along the dimensions they were. We show how this interface can be used to discover trends in a text corpus, such as free-text responses to a survey.

Author Keywords

n-dimensional visualization, common sense, svd, natural language processing

ACM Classification Keywords

I.6.9 Simulation, Modeling, and Visualization: Visualization

INTRODUCTION

Language and language understanding plays a large role in the world of human-computer interaction. Users express their opinions en masse on surveys, in forums, and in dialogue systems, creating a need for systems which can help others visualize and understand the meaning of large collections of such data. When working with the semantics of natural language data, we often need to make sense of data that can be measured in many different dimensions – thousands of dimensions or more. This leaves two related problems: how to express the data in such a way that a computer can make sense of it, and how to further generalize the data so that a human can understand and work with it.

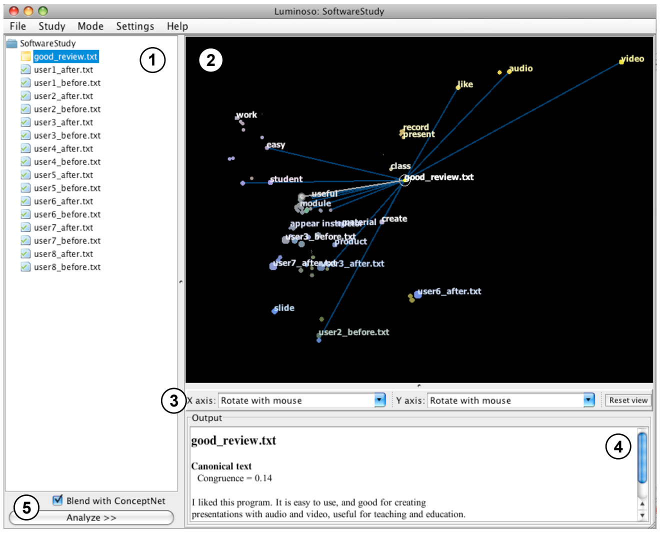
A straightforward computational way to model word co-occurrence among a corpus of documents, for example, is the "bag of words" model, where each document is described with the number of times each word occurs in it. This can easily create a feature space of tens of thousands of features, one for each word that appears in the corpus. In our work, we tend to use not just a bag of words for our semantic models; we also include background common sense knowledge from ConceptNet to provide the models with more "intuition" [7]. This, of course, makes the size of the feature space even larger. A common next step is to use dimensionality reduction to reduce the size of the feature space. Using an algorithm such as truncated SVD reduces the high-dimensional data to a vector space with many fewer dimensions, so that perhaps only 20 or 50 dimensions are necessary to represent the structure of the data.

These vectors, with 20 dimensions or so, are much easier to work with and compare to each other, and they make generalizations that give them more representational power than the original vector space. This is the core idea behind latent semantic analysis (LSA). When a semantic network of background common-sense knowledge is added in, it is also the idea behind AnalogySpace [8], a representation discussed further in the referenced paper.

Luminoso is an interactive application that aids a researcher in exploring these semantic spaces in a way that is intuitive for discovering semantic patterns from the dimensionality-reduced data. It enables them to create a vector space from a folder of input documents, using either an AnalogySpace-based model or a plain bag-of-words model, and then to explore that space interactively on a two-dimensional computer screen. The goal is to help the researcher understand their data by exploring this space, using an intuitive mouse operation we refer to as *grabbing*, which simultaneously lets them visualize the semantic neighborhood of the grabbed data point and use that point to N-dimensionally rotate their viewpoint. Other features of the interface help the user understand the space better, such as by using vectors with known semantics as "signposts".

Using Luminoso is a form of data mining that focuses on interactive exploration of the data. The importance of user participation in data mining has been observed by others [2], because an unsupervised algorithm to detect correlations in data will tend to find correlations that are spurious and irrelevant. An involved user, however, can guide this process toward relevant results by using their intuitive sense of what is interesting.

IUI'10, February 7–10, 2010, Hong Kong, China.
Copyright 2010 ACM 978-1-60558-515-4/10/02...$10.00.



X axis: Rotate with mouse
Y axis: Rotate with mouse

good_review.txt

Canonical text
Congruence = 0.14

I liked this program. It is easy to use, and good for creating presentations with audio and video, useful for teaching and education.

☑ Blend with ConceptNet

Figure 1. The overall interface to Luminoso, with labeled parts: (1) The document pane, allowing documents to be selected and new documents to be added to the study. (2) The viewer pane, providing a two-dimensional view into the SVD space. (3) The axis controls, with which the user can fix one or both of the axes to represent particular directions in the space. (4) The output pane, showing information about the selected point. (5) The "Analyze" button, which runs the SVD and updates the view. The "Blend with ConceptNet" option may be replaced with an interface for blending with any external data set in a future version.

A use case for Luminoso that we focus on is to understand large quantities of people's suggestions and feedback at once. Survey forms frequently contain free-response spaces where people can write a paragraph to explain their views, but after a large enough number of people reply to such a survey, the free-response feedback tends to be ignored. Nobody has the time to read it all. By loading that data into Luminoso, however, one can visualize the major clusters of responses, view representative responses from each cluster, and even include known data about emotion or affect as signposts to understand the tone of the data.

CREATING THE SPACE

In order to display a representative space with Luminoso, we must first analyze the textual input used to create the space. We do this using a series of techniques designed to find patterns in natural language data – including patterns that appear within the input data, that come from a corpus of background knowledge, and that become apparent in the conjunction of both. We can use these techniques to draw general conclusions about the meaning of the data, cluster information in a variety of semantically informed ways, and make inferences across different types of information.

Natural language is a mode of input that can be handled particularly well by our techniques of common sense reasoning. We amplify the power of LSA, which is based only on the co-occurrence of words among the input documents, by including additional information about the semantic connections between words from ConceptNet. The additional knowledge this provides can help to better organize the words and phrases that appear in the input documents into a semantic space. It can recognize when two different words are semantically close to each other, such as "audio" and "video", even when this is not apparent from the distribution of word occurrences in the documents. It can also distinguish words that appear in different topic areas that exist independently of the input data, such as "action verbs", "household items", "computer terminology", and "things people don't want". This kind of information gives the vector space more power to represent the rough meaning of a document.

Applying common sense

ConceptNet [6] is a semantic network created using the information collected by the collaborative Open Mind Common Sense project. Using a representation that expresses knowledge as relations between words and short phrases, it describes the meanings of the words people use in terms of other words. The information contained in ConceptNet includes relations between everyday objects ("Books are used for reading."), information on people's priorities and goals ("People want to be respected."), and affectual information ("Arguments make people angry.").

AnalogySpace [8] refers to the technique of reasoning over such a semantic network by representing it as a matrix and performing singular value decomposition on it. Information in ConceptNet can easily be transformed into a matrix representation that relates its nodes (concepts such as "dog" or ”taking pictures") to their neighboring edges (features such as "…has four legs" and "…is used for enjoyment"). Singular value decomposition expresses these concepts and features in terms of a core set of *axes*, or principal components, that are selected by the algorithm to represent the most variance in the data. The effect is to summarize the provided common-sense knowledge in terms of its large-scale patterns, using moderate-sized vectors (typically 50 to 100 dimensions) to represent each concept and each feature.

INTERACTING WITH LUMINOSO

The first step in interacting with Luminoso is to load the input documents. The container that holds documents and their analysis is called a *study*. The user can use the document tree to add documents to analyze (or they can use their operating system to drop documents into the folder representing the study). One or more of these documents can be marked as "canonical", which highlights it in the tree and makes it stand out in the interface, with effects that will be described later.

The user can choose whether to blend the data with ConceptNet in order to provide background information about semantics. Once the input is set up, the "Analyze" button creates the blend (if necessary), performs the SVD, and displays the results in the viewer window.

The "grabbing" operation

Many of the ways that a user interacts with the Luminoso visualization is centered around an operation we call "grabbing". This operation combines the action of selecting or

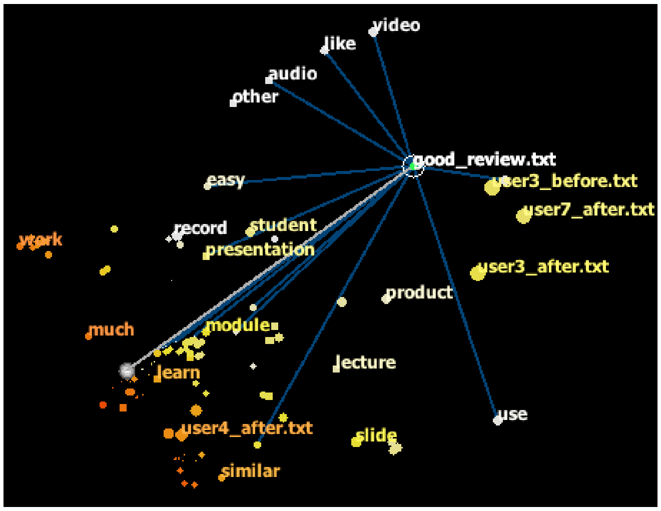


Figure 2. The selected point is a canonical document representing the expected content of a good review. The gray line connecting the point to the origin is always shown, as a reference for comparing with other points.

focusing on a point with the ability to use that point to shift one's perspective in the N-dimensional space. The key idea is that a point you grab is a point you want to understand more about. Thus, not only does "grabbing" a point make information specific to that point visible in the interface, it also lets you find a projection of the data in which you can see the point and its related data better, following the visualization principle of focus-plus-context.

Grabbing a concept and dragging it uses that concept's position to change the projection of the data, as described further on page . It also displays information about the point being grabbed, such as the document's text if it represents a document, and blue lines representing the other terms and documents that the selected point is connected to.

While the point is being held, Luminoso changes the colors of all points to show their amount of correlation with the grabbed point. The colors represent the range of cosine similarity, from -1.0 to 1.0, on a "heat" scale: the most related points glow white or yellow, while unrelated points are a more neutral orange, and diametrically opposed points appear in dark red, as shown in Figure 2.

One way to benefit from this kind of interaction is to add documents with known semantic values into the space. These documents, known as canonical documents, can act as "signposts" when exploring the space. In the interface, a gray axis connects the canonical document point to the origin, informing you that a particular direction corresponds to a particular meaning, no matter what projection you are looking at at the time. When working with a data set of software reviews, for example, a useful canonical document to create is one you construct to represent an idealized good review. Having such a document with a known semantic value, other documents can then be compared to it – either by their location, using the similarity color scale, or by actually locking the X or Y axis of the view on a canonical document.

Congruence

A common use for a canonical document is to test whether the input documents generally "agree" with it semantically. As a way of assisting experimentation, the interface presents a statistic called *congruence* in the info pane when a canonical document point is grabbed. Congruence measures how much that canonical document aligns with the other documents in the study, which can also be seen as describing whether that document is typical or atypical among the input data. This value can be compared between different runs of Luminoso or between different canonical documents.

The congruence of a document is calculated by comparing the distribution of cosine similarities between that document and all others, with the distribution of cosine similarities between all pairs of documents. The congruence is expressed as a Z-value (the difference in means over the standard error), so that it is scale-free.

GRAPHICALLY REPRESENTING N-DIMENSIONAL DATA

After using SVD to describe the data according to its principal components, one is left with vectors with a moderate number of dimensions. At this point, it is Luminoso's job to present this data understandably on a two-dimensional computer screen, so that the researcher can explore the resulting space, see whether it captures the patterns in the input data that it was intended to capture, and discover new patterns along the way.

The data can be represented as a sort of N-dimensional scatter plot. Each word, phrase, common-sense feature, or document in the input corresponds to a point in this space, which will use Luminoso to explore.

At any given time, Luminoso will project all the points in the N-dimensional space onto a two-dimensional plane, which the user can see a part of in a window on their computer screen. The user can change their viewport into this plane much like they would change their viewport in another 2-D interface such as Google Maps: the user can pan by dragging the right mouse button, or zoom using the mouse wheel or a laptop's equivalent "scrolling" gesture.

We represent each point as a small circle, at the appropriate location in the 2-D projection. The size of each circle increases with the number of times the item appears in the input, in order to draw attention to more significant inputs. Every point has a text label, describing a concept, a common-sense feature, or the name of a document, but not all of these labels can be displayed at once – the result would be incomprehensible clutter as many thousands of labels competed for screen space. Instead, only a subset of the labels are shown, determined interactively using the mouse pointer. The labels are chosen so as to set a maximum on the density of labels per unit of screen space. Additional points in a "full" area of the screen go unlabeled. The maximum density of labels decreases with the square of the distance from the mouse pointer. The effect is that, if the user wants to see the label of a point that is currently unlabeled, they can do so by moving the mouse closer to it.

Pressing the left mouse button will select the nearest point and “grab” it, which makes a number of useful things happen, one of which is that the user can use the grabbed point as a handle with which they can transform their view of the N-dimensional space. When the user grabs and drags a point, the view transforms (by stretching and rotating) in such a way that the point’s projection onto the screen follows the mouse pointer, while the origin stays in the same place. The following section describes how this occurs.

Transforming the view

When using Luminoso, it is important to be able to fluidly change the projection of the points onto the screen, in order to see the structure of the data in many dimensions. Frequently, the user is looking for something specific – she wants to place a particular point in a particular location on the screen, and then examine where other points fall around it. For this reason, we allow the user to determine the projection by grabbing points and putting them in particular places on the screen.

The current projection can be described by two vectors in N-dimensional space: a vector $\mathbf{x}$ that represents the current X-axis, and a vector $\mathbf{y}$ that represents the current Y-axis. We add or subtract a small multiple of the grabbed point’s vector from the X and Y vectors, which has the effect of stretching and rotating the space until the point is in the desired place. We also include an option to gradually re-orthogonalize the vectors over time, making these transformations into true rotations that preserve distances and angles.

Related work

Duffin and Barrett [4] describe an interface for rotating a projection, which differs from ours in that the user rotates the space by clicking and dragging a representation of an axis, instead of by clicking and dragging points in the space.

Buja et al. [1] describe the theory of projecting N-dimensional data onto a 2-dimensional view. This paper is largely concerned with creating “tours” of the space, or animations that trace a path between all possible projections of the N axes, but also mentions the ability to rotate particular axes using the “spider” interface.

Buja et al.’s paper provides a survey of existing software for multi-dimensional visualizations, such as GGobi [3], which performs singular value decomposition and allows visualizing the space using 2-D tours and spiders.

APPLICATIONS

Increasingly, the commercial world has become interested in computational linguistics as a way to solve the problem of understanding customer feedback. Focus groups, consumer surveys, and other opportunities to communicate with customers often involve understanding their spoken or written text and “reading between the lines” to understand the patterns. Thus blending common sense with customer-generated free text can often yield insights that normal statistics miss.

The OMCS project worked with a large software company to analyze the data from their user tests [5]. In addition to rating various aspects of the software on a scale from 1 to 7, the users provided short-answer responses to various questions about their perception of the software. This free text data was considerably more expressive and informative than the numeric ratings, but the data was difficult to analyze automatically by computer. Blending with ConceptNet and exploring the results using Luminoso helped to draw general conclusions from the sparse data contained in the free text.

Luminoso can also be used to help users create and develop specialized semantic networks, such as biological or medical information resources, by providing a visualizer which shows the layout, focus, and coverage of the developing resource. As a semantic network gets larger, a multidimensional projection such as that provided by Luminoso becomes a crucial part of visualizing the data.

REFERENCES

1. A. Buja, D. Cook, D. Asimov, and C. Hurley. Computational methods for high-dimensional rotations in data visualization. In C. R. Rao, editor, *Handbook of statistics: Data mining and data visualization*, pages 391 – 415. Lavoisier, April 2005.

2. A. Ceglar, J. F. Roddick, and P. Calder. Guiding knowledge discovery through interactive data mining. pages 45–87, 2003.

3. D. Cook and D. F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*. Springer, December 2007.

4. K. L. Duffin and W. A. Barrett. Spiders: A new user interface for rotation and visualization of n-dimensional point sets. In *In Proceedings of the Conference on Visualization (Los Alamitos*, pages 205–211. IEEE Computer Society Press, 1994.

5. C. Havasi. *Discovering Semantic Relations Using Singular Value Decomposition Based Techniques*. PhD thesis, Brandeis University, June 2009.

6. C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2007.

7. C. Havasi, R. Speer, J. Pustejovsky, and H. Lieberman. Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent Systems*, July 2009.

8. R. Speer, C. Havasi, and H. Lieberman. AnalogySpace: Reducing the dimensionality of common sense knowledge. *Proceedings of AAAI 2008*, October 2008.