There are things that five year-old humans can do easily, while supercomputers cannot do. For example, open-ended tasks dealing with unstructured data such as the natural language or the vision images are challenging to fully automate. I have researched ways that computers can support humans to scale up humans' ability to understand large text corpus using interactive information visualization as a medium to deliver the statistical results to human. However, the bag-of-words model assumption averages the semantic signal such that only the strong topic signal survives. To overcome this limitation, I propose the theory and algebraic algorithm of the thinking that can be an reasoning engine for text data and a unifying theory for the artificial general intelligence.

Visual Analytics for Open-ended tasks in Text Mining

Text mining extracts valuable insights from a text corpus. Many problems in text mining such as identifying characteristics of a group of documents, selecting high-quality comments to promote, or describing an image are open-ended tasks. An open-ended task is difficult to scale up using traditional machine learning approach because (1) the output is unstructured, (2) there can be many different but justifiable answers, and (3) the knowledge about the world is required to solve. The visual analytics approach augments and amplifies human's ability by combining information visualization and statistical machine learning.

My research explores the design space for transforming state-of-the-art text mining algorithms into interactive analytics processes using visual representations. To amplify the cognitive ability of the human analyst to manage large data, In my Ph.D. dissertation work, I have developed four techniques that support the visual text analytics process as described below.

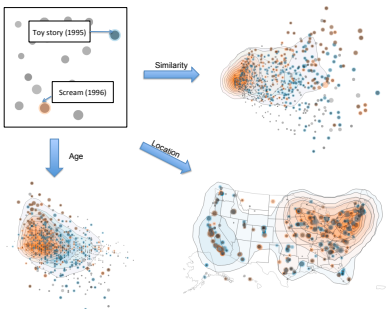

Figure 1: Two movies, *Toy Story* (1995), in blue, and *Scream* (1996), in orange, are compared according to age, location, and similarity criteria of consumers using the ParallelSpaces tool [**park2016parallelspacess**].

ParallelSpaces is a novel method to explore bipartite datasets such as movie ratings data in both feature and similarity dimensions [1]. ParallelSpaces creates side-by-side scatterplots and assigns separate 2D spaces to each class of an entity. Selections in one space are highlighted in the other space, thereby forming visual patterns in the views. For example, the characteristics of the movie can be explained by the people who liked it, and vice versa. Textual dataset with associated meta data such as ratings can be analyzed byvisualizing document term matrix using this technique, where documents clusters can be characterized by words that are frequently used in those clusters.

TopicLens is a Magic Lens-type interaction technique, where docu-

ments under the lens are clustered according to topics in real time [2]. The problem with the previous document galaxy view of the topic modeling is that the number of topics or the level of detail is pre-fixed. TopicLens solves this problem by applying dynamic hierarchical non-negative matrix factorization (NMF). The documents under the lens are sub-clustered by one deeper level of topic, thus revealing underlying subtopics.

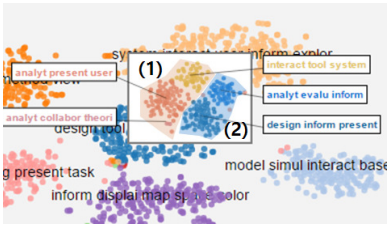

Figure 2: By applying the TopicLens [2] on the clustered papers on visualization, the sub-topics are revealed.

CommentIQ is a comment moderation tool where moderators can adjust model parameters according to the context and goal [3]. Online comments submitted by readers of news articles can provide valuable feedback and critique, personal views and perspectives, and opportunities for discussion. The varying quality of these comments require that publishers remove the low quality ones, but there is also a growing awareness that by identifying and highlighting high quality contributions, the general quality of the community can be raised. Working closely with publishers, moderators, and reporters from the New York Times, the Washington Post, the Baltimore Sun, and the Wall Street Journal, I first created a domain characterization for online comment moderation for news articles and then derived a model for ranking high-quality comments using an annotated comment dataset. Using this model, I designed a web-based visual analytics tool called CommentIQ that generates a visual overview and custom ranked list of comments to be moderated. The full version of the system was evaluated with an extended panel of domain experts, and the feedback strongly supports the utility of the tool for moderating online news comments.

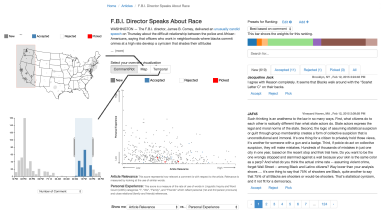

Figure 3: The CommentIQ [3] UI showing toggleable visualizations such as scatterplot, map, and timeline (left) that enable overview and filtering of comments, as well as an adjustable ranking based on various weighted quality criteria (right).

ConceptVector uses word embedding to analyze the text corpus using user generated dictionary [4]. Central to many text analysis methods is the notion of a textual *concept*: a set of semantically related keywords characterizing a specific object, phenomenon, or theme. Textual concepts have potential for characterizing document collections, and can also be constructed once and then shared and reused over and over. I developed a visual analytics system called ConceptVector that guides the user in building, refining, and sharing such concepts and then use them to classify documents. Such concepts can be used as a user-driven feature for the text mining tasks.

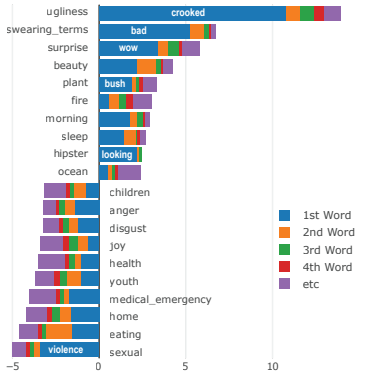

Figure 4: Comparing the tweet messages by 2016 U.S. Presidential election candidates, Donald Trump and Hillary Clinton. The bar chart shows top 10 categories of the words the candidate used more often than the other. ConceptVector [4] is an interactive visual analytics tool for lexicon-based text mining to support creating and applying dictionaries for custom concepts.

Future Research on Artificial General Intelligence

To visualize the text, we need to convert it into the numbers. My previous works rely on the bag-of-words (BOW) assumption, resulting that delicate semantic information is lost. This reduces the productivity and limits scalability to handle large text corpus.

Recent advances in deep learning methods show a possibility to re-

place the traditional statistical machine learning based on the BOW assumption with human-like reasoning engines. When I first encountered the artificial neural net (ANN) back in 2001 to detect parasitic eggs in the microscopic images, I thought it was hopeless compared to other methods such as the Support Vector Machine(SVM). Having taken the neurophysiology course at the medical school during my biomedical master degree, I knew that human brain has to be a simple structure to be built from the genetic information. Jeff Hawkins suggested on his book, *On Intelligence*, that we did not arrange the device in the correct structure, as we have previously done with transistors. Recently, deep learning methods strengthened this view with the successes with Convolutional Nerual Net (CNN) and Recurrent Neural Net (RNN) structures for various tasks. Despite these advances, we do not know yet how we can combine these task-specific models to build an artificial general intelligence.

To tackle these challenges, I propose a unifying theory and an algorithm of thinking. Thinking or intelligent behavior can be represented as a random walk on a probabilistic knowledge graph, where nodes represent concepts, relations or actions available to agents. Edge weights between two nodes are the probability of the next jump. This probability is also conditioned on the context meaning that the optimal thinking or behavior depends on the context. For example, if we are looking at the computer, *apple* might mean the Apple Computer, but when we are eating dinner, apple means the fruit *apple*. For the simple case, where we can assume the context is the previous node, the number of parameters required to build this model are all the probability of jump between all nodes conditioned on the previous nodes. This parameter space is $O(n^3)$, where n represents the number of concepts in the world.

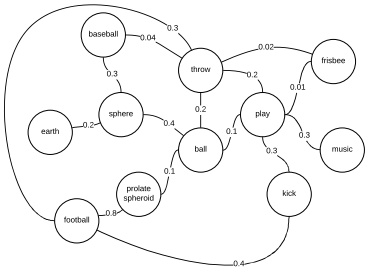

Figure 5: Thinking can be mathematically represented as random walk in the probabilistic knowledge graph, where each node represents sensory input/concept/action. The edge weights represents the strength of association. The edge weights changes dynamically according to the context. To make an intelligent agent, we need at least $O(n^3)$ parameters, where n represents the number of the available concepts/actions.

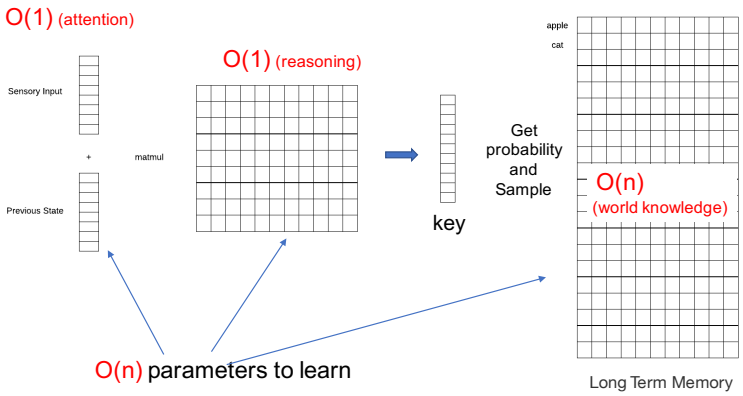

Figure 6: The proposed algorithm for the thinking theory that compresses the $O(n^3)$ parameters for the probabilistic knowledge graph into $O(n)$ parameters. This model is inspired by the recent advances in the attention/memory/RNN.

This ideal model can be algebraically approximated using the al-

gorithm inspired by the biological intelligent agents with neural circuitry. My algorithm is composed of the attention matrix, $O(1)$, on the sensory and short-term memory, the reasoning matrix, $O(1)$, and the long-term memory, $O(n)$, which represents the concepts as the rows of vectors as shown in the Figure 6.

The previous nodes and sensory inputs are selectively concatenated with the attention matrix. The resulting vector is multiplied with the reasoning matrix to produce the key vector, which is then multiplied with the long-term memory vector to assign a probability for every concept. A node is sampled based on this probability. The resulting node is used to generate thinking for one time-step and added to short-term memory. The beauty of this neural algorithm is that it compresses the $O(n^3)$ parameter space of thinking model into the $O(n)$ parameters to make learning tractable for the biological intelligent agents.

References

[1] **Park, Deokgun**, Jungu Choi, and Niklas Elmqvist. Parallelspaces: simultaneous exploration of feature and data for hypothesis generation. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 1437–1445. IEEE, 2016.

[2] Minjeong Kim, Kyeongpil Kang, **Park, Deokgun**, Jaegul Choo, and Niklas Elmqvist. Topiclens: efficient multi-level visual topic exploration of large-scale document collections. *IEEE transactions on visualization and computer graphics*, (Proc. VAST 2016), 23(1):151–160, 2017.

[3] **Park, Deokgun**, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1114–1125. ACM, 2016, **Best Paper Honorable Mention Award**.

[4] **Park, Deokgun**, Seungyeon Kim, Jurim Lee, Jaegul Choo, Niklas Diakopoulos, and Niklas Elmqvist. Conceptvector: text visual analytics via interactive lexicon building using word. *IEEE transactions on visualization and computer graphics*, (Proc. VAST 2017), to appear, 2018.