A Tool for Visualizing and Exploring Relationships among Cancer-Related Patents

Matthew Whitehead and **Daniel K. N. Johnson**
Colorado College
14 E. Cache la Poudre St.
Colorado Springs, CO 80919

Abstract

In this paper, we present a prototype for an online interactive visualization tool for analyzing the semantic proximities of US patent documents that are related to cancer treatments. This tool allows the user to perform keyword searches and then presents visualizations of sets of relevant patent documents clustered by semantic similarity. Semantic similarity is calculated using a combination of word embeddings obtained using the skip-gram algorithm and the t-SNE dimensionality reduction algorithm. The user may then select individual points in the cluster to view more detailed patent information. This process allows the user to explore the connections between related patents and see more general trends in the semantic shape of the technological space. It is our hope that this tool may serve as one of the starting points for data analysis leading to future innovative approaches to cancer treatment.

Introduction

The US Patent and Trademark Office (USPTO) typically grants several thousand new patents per week. This pace of innovation leads to many exciting inventions and discoveries, but the sheer quantity of grants can make navigating the patent landscape daunting. To be successful, inventors must understand how and where their inventions fit in the patent universe in order to ensure that their patent applications eventually end in granted patents that do not pose high risks of future litigation.

In this work, we present a patent data visualization and analysis tool that we hope will increase a user's understanding of the existing patents and how they relate to one another. In particular, we are interested in US patents relating to cancer therapies and our system uses this subset of patents as a prototype dataset. It is our hope that researchers, inventors, policymakers, funding agents, and the general public may be able to use our visualization tool to find unexpected connections between existing cancer-related patents and that these connections may lead to greater innovation and faster progress in this important domain. Visualization tools created with similar goals in mind for web-based searches are described in (Turetken and Sharda 2005).

Our visualization tool is based on building semantic document vector embeddings that are formed using word embeddings from the skip-gram algorithm, which is commonly used in many natural language processing applications. Since these embeddings have been shown to store some semantic word content, they allow our system to group together meaningfully related documents as is commonly done in information retrieval systems. We then use the t-distributed stochastic neighbor embedding (t-SNE) algorithm to reduce the dimensionality of our document embeddings to produce visually appealing representations of the sets of related documents. We make our system interactive by allowing the user to click and select individual documents from within the presented cluster to show patent metadata for further exploration.

By investigating patents both in close proximity to one another and patents that are separated by greater distances, the user is then able to gain some insight into how the technologies and inventions related to the chosen aspect of cancer research are grouped together or spread out over the semantic embedding space. Widely spread clusters of points may signal that there are opportunities for further innovation to fill in the gaps. Tightly packed groups of patents may show that there is a great level of activity surrounding a particular research idea or cancer therapy, perhaps leading funding agents to determine that more resources are required to be devoted to this trending direction of innovation. On the other hand, an individual inventor may see tightly packed clusters of patents as an indication that a particular research direction is crowded, highly competitive, and could possibly lead to litigation risk.

Throughout the rest of this paper, we describe the details of constructing our system. We also present several example visualizations and a short example usage case. Finally, since our tool is currently a prototype and limited in several important ways, we end with a discussion on future improvements and how they might be implemented.

System Description

Our visualization tool produces patent document clusterings when the user searches for keywords. This section describes the process used to build the system including the following:

1. Obtaining raw datasets

2. Preprocessing and cleaning datasets to remove metadata and unwanted characters
3. Generating word embeddings for all words in the lexicon
4. Using word embeddings to generate patent document embeddings
5. Precomputing search results for cancer-related keywords
6. Performing t-SNE dimensionality reductions on each set of search results
7. Generating PNG images for each of the t-SNE clusterings
8. Creating a simple user interface allowing for user interactivity to explore the t-SNE clusterings by use of keyword search and selection of individual patents

The following subsections describe each of these steps in more detail.

Dataset

Our system uses two datasets from the USPTO (USPTO 2016b). The first is a collection of full-text patent documents and associated XML metadata. These documents are available as bulk downloads from the USPTO website with each archive file containing all patents granted within a single week period. We downloaded all patents from January 5, 2010 to March 17, 2015. This produced a collection of 1,337,682 patent documents. Using a larger dataset is possible, of course, but our prototype uses this limited collection in the interest of reducing computational time and storage resources. As discussed below, word embeddings using the skip-gram algorithm continue to improve in quality as more data is processed, so a larger system indexing all US patents would hopefully have even more accurate word embeddings.

The second dataset is a list of 269,354 cancer-related patents offered by the USPTO to support its Cancer Moonshot Challenge (USPTO 2016a), which is part of the larger Cancer Moonshot program (Moonshot 2016) introduced by President Obama in his 2016 State of Union address. The goal of the Cancer Moonshot program is to invest in and accelerate cancer research. We use this dataset to guide user searches and to reduce the number of returned results so that the resulting visualization will be more targeted and allow for easier user interaction. We use the 32,611 patents in this set that fall within the time-frame of our full-text patent dataset described above.

Dataset Preprocessing

The bulk, full-text patent archives contain XML files for each of the granted patents for one week. We preprocessed these files so that their content could be easily used in subsequent parts of the system. First, we extracted each of the patent descriptions from their XML markups. We then converted all the text to lowercase and stripped any non-alphabetic characters with the exception of space characters. Each cleaned document was then concatenated together into a single file. This final file is 47GB in size.

The Cancer Moonshot Challenge dataset contains only metadata about cancer-related patents, so the preprocessing step only worked with patent titles. These titles were converted to lowercase and non-alphabetic characters were removed.

Word and Document Embeddings

Our system uses words embeddings to help find meaningful relationships between cancer-related patent documents. Word embeddings are vector representations of words that can capture meanings in numeric ways. These representations can then provide direct comparisons of the semantic relationships between individual words, and more importantly for the current application, between full documents. Word embeddings are obtained from large corpora of plain-text data by finding patterns in existing written language, and then extracting those patterns in a compressed way as numeric semantic features.

There are many common ways to perform this type of feature extraction including latent semantic analysis (Deerwester et al. 1990), neural-probabilistic language models (Bengio et al. 2003), deep learning models (Collobert and Weston 2008), and the skip-gram algorithm with negative sampling (Mikolov et al. 2013). Since the skip-gram model has been shown to have state-of-the-art performance on a number of prominent natural language processing tasks, we use it for this work.

The skip-gram algorithm generates word embeddings by using artificial neural networks to process words along with their surrounding context. Skip-gram models are trained by processing large volumes of text data and then adjusting their parameters to minimize context prediction error. Once the model is trained, word embeddings are obtained by capturing the model's internal activations in its projection layer when the model is given as input a single, encoded word. Since word context prediction is the goal during training, the activations become semantically meaningful in order to perform better predictions, given the assumption that semantically related words co-occur more frequently than unrelated words. Figure 1 shows a diagram of the model with the projection layer representing the internal network activations.

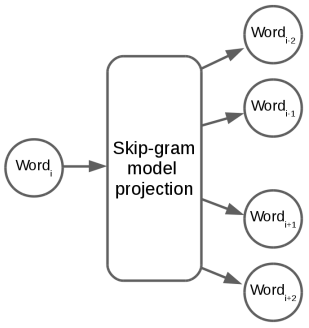

Figure 1: Diagram of the skip-gram model.

For this work, we use the word2vec implementation (Word2Vec 2016) of the skip-gram algorithm. We generated word embedding vectors of length 300 by training on the

entire patent dataset that we gathered, including patents unrelated to cancer therapies.

We use patent document embeddings in order to group together semantically related documents. The document embeddings attempt to capture the combined semantic information contained within each patent document. Embedding documents in a space based on the word embeddings allows us to directly compare not only documents to other documents, but also documents directly to words. Our method for creating document embeddings is simple: use the centroid of all the word embeddings vectors for all the words contained in the document. Other methods for creating more meaningful document embeddings have been proposed, such as in (Le and Mikolov 2014), so this is one aspect of our tool that could be upgraded on the next iteration.

Dimensionality Reduction with t-SNE

After generating document embeddings for each of the patent documents in our dataset, we then produced cluster images for possible cancer-related keyword searches. These images are shown to the user when searches are performed.

We formed the clustering images using the t-distributed stochastic neighbor embedding (t-SNE) algorithm (van der Maaten and Hinton 2008). The t-SNE algorithm creates probability distributions for pairs of datapoints in both the original high-dimensional space and the low-dimensional space of the mapping. These distributions contain higher probabilities for pairs of datapoints that are more similar and lower probabilities for pairs of points that are less similar. The low-dimensional probability distribution is formed while attempting to minimize the Kullback-Leibler divergence using gradient descent. The resulting mapped datapoints can then be visualized in a 2-dimensional scatterplot. The algorithm has been shown to have good performance on a number of different datasets for high-dimensional visualization.

To select the set of keywords to use for the t-SNE image generation, we used the titles of all cancer-related patents in the USPTO Cancer Moonshot Challenge dataset. For every word present in any of the titles, we collected all documents sharing that keyword and then executed the t-SNE dimensionality reduction on that set. Keywords present in fewer than ten patent titles were excluded. For clarity of data visualization, returned sets were also capped at 200 documents.

After completing the t-SNE execution on each keyword document subset, we then generated a PNG scatterplot image showing the documents embedded in 2-dimensional space. For our chosen dataset, this resulted in about 2400 pre-generated images. These images are pre-generated so that they may be presented without much delay when the user is interactively using the system. Search queries containing multiple words are not yet supported in our prototype because of the additional resources required, but it is a feature to be added in the near future.

Interactive Visualizations

The frontend is written in javascript and uses the JQuery library (JQuery 2016). The pre-generated clustering graphs and their metadata files are stored on the server and indexed by keyword. When the user enters a search query, then the corresponding keyword scatterplot is displayed.

Once the graphs are displayed, the user is then able to explore the results by selecting individual patent documents. Clicking a patent's corresponding dot in the scatterplot displays the patent's title and also brings up a link to the patent's full text. By selecting patent document dots in close proximity to one another, a user may explore a subset of the semantic space covered by those documents.

Since the clustering images are precomputed, the responsiveness of the system is very high. Currently, when the frontend system first loads, the metadata for all the documents is downloaded and prepared for display. This download is currently several megabytes of data adding a slight delay (depending on the internet connection of the connected user) to the interface's start-up time, but in the future we plan on making this metadata load on demand to speed up initialization.

Discussion

Many patent clusters show unsurprising groupings of closely related documents, but it is our hope that others will shed light on unexpected relationships that are meaningful. For example, if a patented process attempting to solve a difficult problem did not succeed, then this tool may help with the search for the next solution. This tool can also guide the search for neighboring technologies or techniques, either to secure against litigation or infringement concerns, or to find alternative/substitute approaches to similar issues.

Our visualizations also provide a projected outline of the documents under consideration, drawing attention toward the edges, toward the holes, or toward the clusters. Where might investors or translational scientists find the most creative patents, the ones with the greatest distance to other patents? Where might they find the edges of the discipline, the opportunity for advancement without competitors or previous knowledge? How might we extend our knowledge by pushing the edges of the graph outward? Where should we focus our energies on backfilling gaps in our research knowledge?

As an example usage of our system, consider Figure 2 that shows two screenshots of a single search for the keyword "cell." Each image shows the same set of results, but with a different patent selected. Both selected patents are nearby to one another, so they should be semantically related (or at least more closely semantically related than other patents in our chosen dataset). The left image shows US patent 7862814, "Method of inhibiting the proliferation of B cell cancers using TACI-immunoglobulin fusion proteins," and the right image shows US patent 8216576, "Method for inhibiting binding to B-cell receptor." A user may then investigate both of these patents simultaneously and perhaps find important similarities and differences in both technique and patent focus, allowing for new insights and directions of innovation.

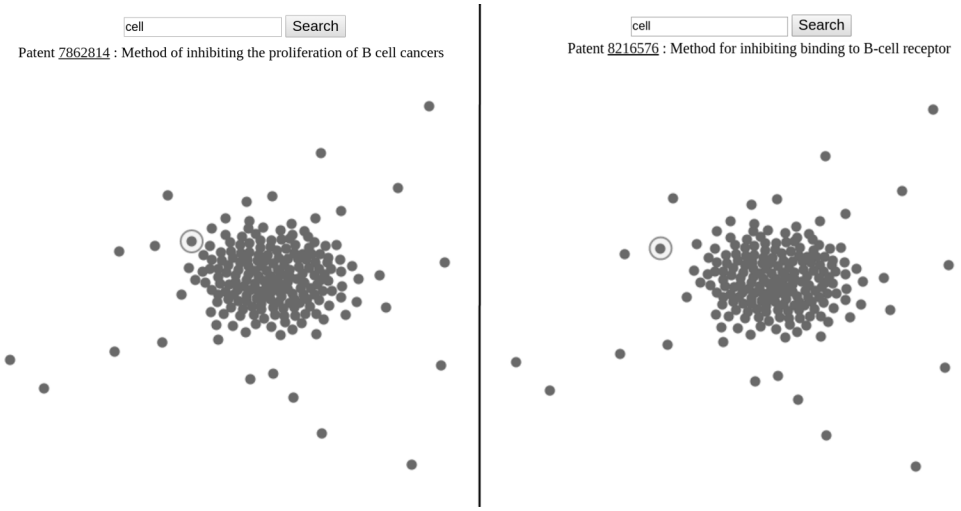

Figure 2: Two screenshots of a user search for "cell". Each shows a different document selected (with selection indicated by a circle around one patent dot), but both are in close proximity to one another. Both patents (US7862814 and US8216576) concern B cell cancers and, more specifically, inhibitions of both receptor bindings and cancerous cell proliferation.

Future System Improvements

Since our existing visualization tool is currently an example prototype of an emerging application, it is limited in several important ways. Our working dataset only contains patents from about five years' worth of innovation. Filling out the system with all available patents would not require a change in any of the preprocessing steps other than more time and computational resources.

Keyword searches are limited to single-word queries. This, again, was done in the interest of lowering computational costs for getting an initial prototype system up and running quickly so that it could be used to guide future improvements in the user experience. For the system to support longer queries but still remain highly interactive with no delays between queries, the backend infrastructure will need to be upgraded. Since the t-SNE algorithm is fairly computationally intense, all visualizations must be pre-generated. Moving to multiple-word queries will then involve either pre-generating large quantities of graphs with the various combinations of keywords or finding a faster way to compute t-SNE reductions on-the-fly. Implementing a caching mechanism would likely help speed up this second possibility since many shorter queries would likely be reiterations of previously seen queries.

Acknowledgments

We gratefully acknowledge the NVIDIA Corporation for the donation of the Titan X GPU used to support this research.

References

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3:1137–1155.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In Cohen, W. W.; McCallum, A.; and Roweis, S. T., eds., *ICML Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, 160–167. ACM.

Deerwester, S.; Dumais, S. T.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

JQuery. 2016. Jquery. https://jquery.com/. Accessed: 2016-09-10.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. *CoRR* abs/1405.4053.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality.

Moonshot, C. 2016. Cancer moonshot. http://www.whitehouse.gov/CancerMoonshot. Accessed: 2016-09-10.

Turetken, O., and Sharda, R. 2005. Clustering-based visual interfaces for presentation of web search results: An empirical investigation. *Information Systems Frontiers* 7(3):273–297.

USPTO. 2016a. Cancer moonshot challenge. https://www.challenge.gov/challenge/uspto-cancer-moonshot-challenge/. Accessed: 2016-09-10.

USPTO. 2016b. United states patent and trademark office. http://www.uspto.gov. Accessed: 2016-09-10.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Word2Vec. 2016. Word2vec. https://code.google.com/archive/p/word2vec/. Accessed: 2016-09-10.