A Content-Based Recommendation System for Online Communities at High Scale

Robert Elwell
Wikia, Inc.
robert@wikia-inc.com

Kevin Cooney
Wikia, Inc.
kevin@wikia-inc.com

Tristan Chong
Wikia, Inc.
tristan@wikia-inc.com

Chris Fife
Wikia, Inc.
cfife@wikia-inc.com

ABSTRACT

In this paper, we outline a large-scale recommendation system capable of suggesting online communities that may be interesting to visitors of a given community. Recommendations use spatial calculations computed on a matrix of latent topics derived from dimensionality reduction against a sparse set of deep natural language features. We extract the features for this algorithm using a distributed text processing pipeline managed in the cloud. Experimental testing using a random sampling of users shows a statistically significant increase in engagement against the current baseline for wikis recommended using this system, as well against an information retrieval-based approach.

Categories and Subject Descriptors

I.2.7 [**Natural Langauge Processing**]: Text Analysis; H.3.3 [**Information Storage and Retrieval**]: Selection Process; H.3.4 [**Systems and Software**]: Distributed Systems

Keywords

Natural Language Processing, Recommendation Systems, Cloud Computing, Web Scaling

1. INTRODUCTION

Dynamically serving content-based recommendations is an important aspect of servicing communities at high scale. Recommending contextually similar communities along the long tail should increase engagement, improve discovery, and provide a more desirable user experience. We service more engaging recommendations to our top 5,000 English communities, which represent two-thirds of our English content, by using cues from from deep natural language features extracted from the content hosted on each community.

1.1 Background

Wikia uses the MediaWiki platform to provide hosting for wiki-based communities of any subject matter. These communities are individually referred to as wikias. Over 400,000 communities focused in the areas of entertainment, gaming, and lifestyle content are hosted on this platform, in over 70 languages. These communities are responsible for over 30 million pages of content, over half of which is in English.

Advertising is Wikia's core source of revenue. The platform hosts over 2 billion monthly pageviews. Wikia is a Quantcast Top 25 company.

Recommendations are presently provided in a module page entitled "Around the Wikia Network". In the past, these recommendations have been powered by presenting a randomized grouping of wikias belonging to the same vertical as the present wikia in question. Wikias to be presented in this module are hand-selected as being representative of that specific vertical. Wikia classifies its communities into three main verticals: Entertainment, Gaming, and Lifestyle.

1.2 Increasing User Engagement

User engagement is important to Wikia for two reasons. It is primarily beneficial to users to interact with content that is important to them. As part of Wikia's business interest, user engagement is interpreted as a predictor of pageviews. Consequently, we assume pageviews are a predictor of revenue via ad impressions and clicks. We define user engagement as average pageviews per session.

We assert that user engagement can be improved on Wikia by improving the wiki recommendations in the existing "Around Wikia's Network" module. Furthermore, we posit that wikias that are more relevant to the content of the community in question will serve as a better data source for recommendations than randomized hand-curated communities belonging to the same vertical.

2. IMPLEMENTATION

In order to provide content-oriented recommendations, we use deep natural language features extracted from our text processing pipeline. These features serve as the input for topic modeling, which serves to reduce dimensionality and identify affinities between sets of words. Spatial computations are then used on the topic values of each wiki to determine similarity.

Spatial similarity against topic models has been shown to serve as an effective means of improving recommendation quality. The unsupervised nature of the learning model we

choose has been shown to be particularly effective in identifying similarities amongst documents for large-scale social platforms. Concrete instances based on such a platform can vary in their subject matter, and users are responsible for generating their own ontologies. This makes it an especially difficult problem for supervised or rule-based approaches. [2]

2.1 Text Processing Pipeline

Wikia's text processing pipeline employs Stanford CoreNLP to transform raw text into XML files containing deep natural language data.[4] The assets of these XML files include part-of-speech-tags, syntactic parses, dependency parses, coreference-resolved named entities, and per-sentence sentiment. Data from these XML files are extracted and used as features for a topic model, in the case of this recommendation system. [15] [14] [11] [3] [9] [7] [6] [8] [12]

Wikitext for each article is parsed into HTML upon each edit. Upon creation or update, Wikia's search indexing pipeline indexes data about the article, including its raw text. Raw text is extracted from parsed HTML using a DOM parser. This raw text serves as the data source for the parser.

Data is transported from within Wikia's internal network to be computed on Amazon Web Services. This platform provides APIs and interfaces to elastically scale resources such as storage and computing in an automated fashion. This allows us to initiate, monitor, and manage computations from exclusively within our own network.

We use a modification of the Stanford CoreNLP pipeline that supports the polling of a directory of flat files of raw text, attaching an available thread in a set pool to each file. Our modification allows the inclusion of additional files during parse time. This is important because bootstrapping time for the full suite of features used can take upwards of five minutes on an Amazon Elastic Cloud Computing (EC2) m2.4xlarge instance. A separate script is responsible for polling a queue of compressed batches of files, downloading and decompressing the batch, and adding it to the folder polled by the parser. Completed parses are output as XML and uploaded to Amazon Simple Storage Service (S3).

Upon completion of a parse, events are added to a queue indicating a specific XML file has been added or changed. A data extraction pipeline also hosted on EC2 consumes these events. This pipeline stores pre-cached, denormalized representations of data points extracted from the XML file. This data includes counts of named entities – including pronominal cases of coreference – and counts of specific semantic heads. Named entities are cross-referenced with titles and redirects for that specific wiki for higher relevance to the ontology built by the community. These high-quality named entities and all syntactic heads are aggregated at the wiki level, with counts recorded. These serve as data sources for our topic modeling step.

2.2 Topic Modeling

Topic modeling allows us to define a document as a point in N-dimensional Euclidean space. [13] It also allows us to reduce dimensionality and pre-compute relationships between words into probability distributions over their co-occurence in a given set of wikias.

We define a single document in this step as a given wiki, defined by its top 50 most frequent entities and top 50 most frequent semantic heads. The distribution of these values is skewed by its frequency across the wikia. That is, we observe a feature in a document once if it only appears once in that wikia, and twenty times if it appears a total of twenty times across all documents in that wikia.

We build these document instances for 5,000 of the most popular English-language wikias. These wikias represent two thirds of all English-language content. This implies a worst-case problem space of 12.5 million unique features.

We generate a topic model out of these instances via Latent Dirichlet Allocation (LDA). The generative model is built using the `gensim` Python library, which is capable of distributed computation for improved performance over high scale. [10] The desired model is defined as containing 999 topics. We perform statistical pre-processing to determine and remove 300 "stop" features, using a Borda ranking of statistical value and entropy. [16] We also use features of the software library to filter extreme instances while maintaining a minimum set of features. We specify withholding a minimum set of 100,000 features. We are therefore reducing the dimensionality of this pre-processed problem space by two orders of magnitude, compared to its worst case.

The result of running LDA on our learning set is a sparse matrix of values ranged 0 to 1 for 999 different topics. These topics represent probability distributions over groupings of words. They allow us to express a document as a discrete point in hyper-dimensional Euclidean space.

2.3 Similarity by Distance

As we can now express each document as a point in hyper-dimensional space, we can use spatial distance calculations to express the similarity of one document in the set to all other documents in the set. We use the cosine distance metric available in `SciPy` to calculate similarity between each document based on the values of their respective topics. [5] We improve the performance of this calculation against the Cartesian square by indexing nonzero values for a document to determine whether one document shares any nonzero topics with another.

For each document, we calculate distance only against those documents which share at least one nonzero topic value in common. We retain a set of 20 documents in ascending cosine distance as the set of recommendations. These document's IDs are then stored in a location accessible to the Wikia application platform for that article at the time of a request.

2.4 Similarity by Common Subjects

We also explored an alternate method of providing recommendations, using an approach based on information retrieval heuristics. This approach surfaced the top results for a cross-wikia search based on the current wikia's core subject matter.

Wikia's communities are centered around a shared topic of interest. The names of each wikia are generally highly aligned to these topics. However, numerous high-quality wikias use names that do not necessarily indicate the main topic of discussion. For example, The Dr. Who Wikia refers to itself as "Tardis Data Core", and The Lord of the Rings Wikia refers to itself as "One Wiki To Rule Them All".

We employ a heuristic against data used to power Wikia's cross-wikia search functionality to identify the subject of a wikia. These features include a wikia's canonical URL (and any redirect URLs), its site name, its headline, the wikia's

description. For the main page of that wikia, we use the article title, the content, and its HTML title. We also use the most popular titles and categories for that wikia. Each of these features are fields which a candidate subject may appear in.

Short-form data such as titles, categories, and information from the URL is normalized. Long-form data, such as descriptions are parsed, and all noun phrases are extracted and normalized. The normalized phrases are stored in a hash, keying their normalized forms to their original versions. These values serve as candidates for a wikia's subject.

For each candidate phrase, we score 0 or 1 based on set-membership for fields containing a single term and apply a min-max scaled term frequency score for fields with multiple terms. Scores for each candidate phrase are calculated and multiplied by numeric weights for each field. Preference is given to the data from the page title and canonical hostname. We sum the score against each field for each candidate and sort candidates by score descending. We consider the highest-scoring candidate to be the subject of that wikia.

Given a wikia's subject of interest, we perform a cross-wiki search using that term and surface those results – excluding the current wikia – as a candidate recommendation set.

3. METHODOLOGY

In order to test the success of our candidate recommendation systems, we ran A/B tests against four equally sized, randomized groups of users. The groups then received different recommendations in the "Around Wikia's Network" module.

Group A consists of one row of three recommendations based purely on a cross-wikia search against the subject of the current wikia. **Group B** consists of three recommendations based purely on topic distance. **Group C** combines groups A and B by displaying two rows of three recommendations – one based on cross-wikia search, and one based on topic distance. **Group D** serves as our control group, displaying the current treatment of three hand-selected wikias in the same vertical as the current wikia.

1% of traffic to the top 5,000 English wikias was withheld for random assignment one of the above four groups. The A/B test was run on the selected groups for five days between April 14th, 2014 and April 18th, 2014. After the test period completed, we calculated analysis of variance (ANOVA), Bonferroni and Tukey's Honestly Significant Difference (HSD) tests for pairwise comparisons against the four groups.

4. RESULTS

Analysis of variance of the three experimental groups against the control group shows a distinct support for the test group using recommendations solely generated by cosine distance against topics.

As shown in Figure 1, users in Group B, who were shown recommendations based solely on cosine distance against LDA topics, showed a 13% increase in pageviews per session against the control group. For pairwise comparisons, we applied ANOVA to the behavior of each group at an hour-by-hour window. This calculation showed the increase in Group B against the control group (Group D) to be statistically significant ($P < 0.01$).

Figure 1: Results of Experiment on Each Test Group

Group	Visitors	Sessions	Pageviews	Pageviews/Session
A	23,747	4,293	20,741	4.83
B	24,082	4,361	23,825	5.46
C	23,747	3,767	18,790	4.99
D	21,842	3,997	19,287	4.83

We can extrapolate the effect of implementing the recommendations in Group B for our top 5,000 English wikias alone as having a possible impact on the order of hundreds of thousands of additional pageviews per week.

Much of these improvements appear to be from users who visit Wikia regularly. We can see this by cross-referencing pageviews per session by session per user. For visitors with 200 or more sessions during the five-day period of the experiment, we see 17% increase in pageviews per session for Group B, yielding a 70% increase in pageviews when compared to the Group C in total.

Interestingly, for visitors with fewer sessions, search-based recommendations (Group A) seem to have a slight edge. As sessions per user increases, the combined method (Group C) shortly takes precedence over both experimental groups, until Group B overtakes both. A possible explanation for this trend is that users with fewer sessions tend to arrive at Wikia via search engine, and may only be interested in topics highly relevant to their initial topic of interest. High-frequency users may be more willing to explore Wikia's platform, but might lack the ability to discover communities for which they lack awareness. Users in the middle may not yet be aware of other communities that are highly relevant to their pre-existing interests. High-frequency users may already be aware of these communities, but perhaps not those wikias which share certain latent features.

5. FUTURE WORK

Personalizing recommendations based on user path behavior may serve to further improve engagement. Implementing a time-varying user model against our deep language data source could allow for personalized predictions highly aligned with both community content and user behavior. [1] We also plan to apply this model to individual pages as well as video content.

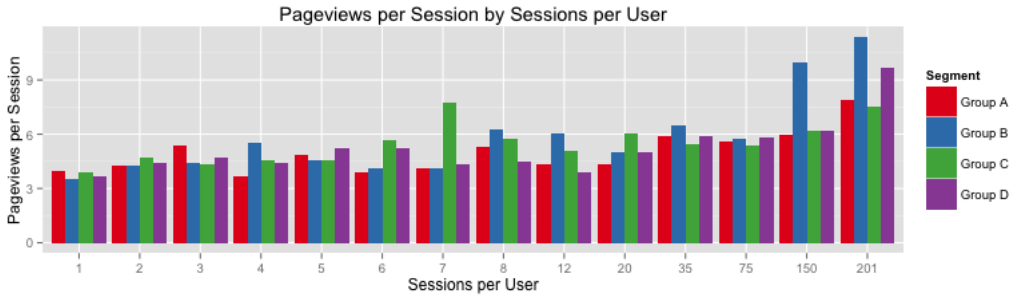
6. CONCLUSION

In this paper, we have outlined a method for using spatial calculations against topics generated from natural language features to identify similar communities of possible interest to a visitor based on subject matter and content. We accomplish this using a distributed, cloud-based pipeline including assets responsible for text parsing, parse data extraction, dimensionality reduction, and spatial calculations. Each of these components has been designed and implemented for processing gigabytes of data while distilling the most contextually valuable information at each stage. Using this approach, we achieve an improvement in engagement which we believe will provide a direct impact to revenue through advertising clicks and impressions.

7. REFERENCES

[1] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J. Smola. Scalable distributed inference of dynamic user interests for

Figure 2: Histogram showing high engagement with NLP recommendations by regular users



behavioral targeting. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 114–122, New York, NY, USA, 2011. ACM.

[2] Konstantinos Christidis, Gregoris Mentzas, and Dimitris Apostolou. Using latent topics to enhance search and recommendation in enterprise social software. *Expert Syst. Appl.*, 39(10):9297–9307, August 2012.

[3] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *IN PROC. INT?L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC*, pages 449–454, 2006.

[4] Robert Elwell, Tristan Chong, and John Kuner. *Taming Text: How to Find, Organize, and Manipulate It, 2nd. Ed*, chapter A High-Scale Deep Learning Pipeline for Identifying Similarities in Online Communities. Manning Publications Co., Greenwich, CT, USA, 2015.

[5] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[6] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, December 2013.

[7] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. The life and death of discourse entities: Identifying singleton mentions. In *HLT-NAACL*, pages 627–633. The Association for Computational Linguistics, 2013.

[10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[11] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing With Compositional Vector Grammars. In *ACL*. 2013.

[12] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October 2013. Association for Computational Linguistics.

[13] M. Steyvers and T. Griffiths. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum, 2007.

[14] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[15] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part–of–speech tagger. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, 2000.

[16] Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, and Lu Sheng Wang. Automatic construction of chinese stop word list. In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, ACOS'06, pages 1009–1014, Stevens Point, Wisconsin, USA, 2006. World Scientific and Engineering Academy and Society (WSEAS).