Semantic Heuristics for A* Pathfinding in a Hyperlinked Reddit Network

Aidan Essig

Source Code

Abstract

This paper investigates the problem of semantic pathfinding in a large subreddit hyperlink network, where nodes represent subreddits and directed edges encode user-posted hyperlinks. Each edge is annotated with an 86-dimensional feature vector capturing linguistic, structural, and psycholinguistic properties of the source post. I propose a heuristic-driven approach using A* search, where the heuristic is based on cosine distances between average feature vectors of subreddits. To evaluate the effectiveness of individual features, I run extensive experiments comparing A* search against breadth-first search (BFS) across randomly selected subreddit pairs. Results show that certain individual features, particularly those related to sentiment, social language, and financial references, lead to significant reductions in node expansions, achieving over 70% savings relative to BFS. Combining high-performing features results in even greater efficiency. These results demonstrate that simple, interpretable heuristics built from post-level properties can make pathfinding in subreddit networks much more efficient.

Introduction

Reddit is home to thousands of distinct communities, or subreddits, each centered around specific interests, cultures, or discourse subjects. Despite their topical boundaries, these communities are not isolated. Users often create hyperlinks in their posts that reference content in other subreddits. These interconnections form a sprawling network of interactions that reflect the thematic, social, and ideological relationships among subreddits.

Inspired by the popular web game WikiRacing, where players race from one Wikipedia article to another by only clicking hyperlinks, I developed Reddit Race, an interactive platform that challenges users to navigate from one subreddit to another using only hyperlinks between posts. The live experience is available at reddit-race.com. While WikiRacing draws on human intuition and the semantic structure of Wikipedia, Reddit Race presents a fundamentally different challenge. Subreddit hyperlinks are often sparse, unpredictable, and shaped by community dynamics rather than formal encyclopedic logic.

To better understand and optimize navigation in this space, I analyze the subreddit hyperlink network extracted by Kumar et al. (2018) (1) in their study of intercommunity conflict on Reddit. This directed network includes over 850,000 edges linking more than 55,000 subreddits, each annotated with an 86-dimensional vector of linguistic and psycholinguistic features. These features capture everything from word count and sentiment to social language and topic-specific LIWC (Linguistic Inquiry and Word Count) categories. The dataset spans Reddit activity from January 2014 to April 2017 and is publicly available through the authors' project page.

The core goal of this project is to investigate whether semantic heuristics, based on subreddit content and language patterns, can meaningfully guide navigation across this network. To do this, I compare standard breadth-first search (BFS) with A* search, using heuristics computed from cosine distances between average post vectors for each subreddit. Through large-scale experiments, I explore the role of individual features and feature combinations in reducing node exploration while preserving path quality.

In what follows, I describe the data and pruning strategy used to construct a more navigable hyperlink network, explain the heuristic-driven methods in detail, present empirical results, and reflect on their implications for search problems.

Background

Pathfinding in a network involves finding a route from a starting node to a goal node, aiming to minimize the number of steps or the overall cost. In unweighted networks like the subreddit hyperlink network in this project, the standard approach is breadth-first search (BFS). BFS guarantees the shortest path in terms of the number of edges, but it explores all nodes at a given depth equally. This can often be inefficient when the network is large or contains meaningful structure that could help guide the search.

To improve efficiency, A* search introduces a heuristic function to estimate how close a node is to the goal. At each step, A* chooses the node with the lowest estimated total cost, which is the sum of the distance from the start ($g(n)$) and the estimated distance to the goal ($h(n)$). When the heuristic is well-designed, A* can find the shortest path while exploring far fewer nodes than BFS.

In this project, the heuristic is based on the semantic similarity between subreddits. Each subreddit is represented by an average feature vector, computed from the text properties of its outgoing posts. These vectors have 86 dimensions, covering attributes like sentiment, readability, and language usage.

To compare how similar two subreddits are, cosine distance is used. First, cosine similarity between two vectors $\vec{a}$ and $\vec{b}$ is defined as:

$$\text{cosine\_similarity}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

This gives a value between 0 and 1, where 1 means the vectors point in the same direction (i.e., high similarity), and lower values indicate increasing dissimilarity. Cosine distance is defined as:

$$\text{cosine\_distance}(\vec{a}, \vec{b}) = 1 - \text{cosine\_similarity}(\vec{a}, \vec{b})$$

We use this cosine distance as the heuristic in A*. If a subreddit has a feature vector that is close to the goal's vector, the heuristic returns a lower value, meaning that node is prioritized in the search. This allows A* to focus on nodes that are semantically similar to the goal, reducing the number of nodes it needs to explore.

Related Work

The closest related work is the project from which I sourced my dataset: Kumar et al. (2018), titled *Community Interaction and Conflict on the Web*, published at the Web Conference (WWW 2018) (1). In that study, the authors analyzed 137,113 hyperlinks between 36,000 Reddit communities to investigate large-scale patterns of intercommunity conflict. They defined and modeled "mobilizations" as cases where a subreddit links to another in a hostile or adversarial context, triggering increased participation from one community in another. Their analysis revealed that a small fraction of communities, just 1%, were responsible for initiating the majority (74%) of all conflict events, and that these conflicts frequently occurred between communities with similar topics but opposing ideologies.

Whereas their goal was to detect and predict toxic cross-community behavior using learned embeddings and a socially primed LSTM model, my work repurposes the same dataset to explore a different question: how semantic similarity between subreddits can be used to guide efficient navigation through the hyperlink network. Rather than focus on social conflict, I approach the network as a semantic space, evaluating how post-level features, such as sentiment, formality, or topicality, can serve as lightweight, interpretable heuristics for pathfinding.

This semantic routing objective is directly implemented and visualized in my interactive platform, reddit-race.com, where users can explore hyperlink paths between communities in real time.

Project Description

This project focuses on semantic pathfinding in a large subreddit hyperlink network using heuristics derived from post-level feature vectors. The overall system is composed of four main steps, described in Algorithm 1 through Algorithm 4.

Algorithm 1: Step 1: Build Pruned Subreddit Graph

```
Count frequency of each (source, target) pair
Filter out links with fewer than 3 occurrences
for each source subreddit do
    Keep top 7 most frequent target subreddits that are also sources
end for
Construct directed graph G from remaining links
```

Algorithm 2: Step 2: Compute Average Subreddit Feature Vectors

```
Load feature vectors from all hyperlinks
Parse and validate 86-dimensional feature vectors
for each source subreddit do
    Compute average feature vector over all its posts
    Store result as mapping: subreddit → avg vector
end for
```

Once the graph G is constructed and feature vectors are averaged per subreddit, I evaluate A* search using cosine distance heuristics based on individual features and feature combinations. A* is benchmarked against BFS.

Algorithm 3: Step 3: Evaluate A* Using Individual Features

```
for each feature index i from 0 to 85 do
    Define heuristic h_i(a, b) = 1 − cos(vec_a[i], vec_b[i])
    for each random subreddit pair (s, t) do
        Run BFS to get path and nodes visited
        Run A* using h_i to get path and nodes visited
    end for
    Record average A* vs BFS performance and savings
end for
```

Algorithm 4: Step 4: Optimize Heuristic via Feature Combinations

```
Select top individual features based on savings (k)
for each combination S of size k do
    Define heuristic h_S(a, b) using cosine distance over S
    for each random subreddit pair (s, t) do
        Run BFS and A* with h_S and record stats
    end for
    Track average visited nodes and % savings
end for
```

Experiments

To evaluate the performance of feature-driven heuristics, I conducted a series of large-scale experiments comparing A* search to BFS over random subreddit pairs. For each trial, both algorithms were used to find paths between 10,000 random (s, t) pairs drawn from the pruned graph G, and the number of nodes visited was recorded.

Individual Feature Trials

Each of the 86 post-level features was evaluated independently by defining a heuristic $h_i(a, b) = 1 - \cos(vec_a[i], vec_b[i])$, where i is the feature index. Results show that many individual features yielded significant reductions in node expansions relative to BFS. The top 10 performing features are listed below as their semantic meaning and node savings:

Feature Name	Savings (%)
Compound_Sentiment	71.74
LIWC_Family	71.55
LIWC_Death	71.51
LIWC_Swear	71.49
LIWC_Religion	71.42
LIWC_Ingest	71.37
LIWC_Friends	71.35
LIWC_Sexual	71.33
LIWC_Assent	71.30
LIWC_Nonflu	71.28

Interestingly, the top-performing individual feature was the compound sentiment score calculated using VADER (Valence Aware Dictionary and sEntiment Reasoner). This feature captures the overall emotional valence of a post by aggregating positive, negative, and neutral sentiment components into a single normalized score. Its strong performance as a heuristic can be attributed to the fact that sentiment often serves as a proxy for ideological tone, emotional framing, and communicative intent, factors that frequently align across topically similar communities. For example, subreddits that express consistently positive or supportive language may cluster around shared identities or interests, while those with negative sentiment may be linked through adversarial or critical commentary. By leveraging sentiment as a semantic signal, the compound score provides an effective and interpretable way to estimate subreddit similarity, especially in the absence of explicit topic overlap.

The other features capture psychologically meaningful aspects of language such as life, religion, and social relationships, all of which appear to influence subreddit-level similarity and community clustering in the network.

Feature Combination Trials

To further improve heuristic quality, I tested combinations of the top 10 highest-performing individual features. For each trial, both algorithms were again tested to find paths between 1,000 random (s, t) pairs drawn from the pruned graph G for each possible combination of the 10 features. The results demonstrate that combining features leads to substantial improvements in efficiency, exceeding 74% node savings relative to BFS. Below are the top-performing combinations by size, with feature indices replaced by their semantic labels:

Top Feature Combination of Size 2

- `[Compound_Sentiment, LIWC_Ingest]` — 73.85% savings

Top Feature Combination of Size 3

- `[Compound_Sentiment, LIWC_Friends, LIWC_Sexual]` — 73.89% savings

Top Feature Combination of Size 4

- `[Compound_Sentiment, LIWC_Swear, LIWC_Ingest, LIWC_Assent]` — 74.12% savings

Top Feature Combination of Size 5

- `[Compound_Sentiment, LIWC_Swear, LIWC_Religion, LIWC_Friends, LIWC_Assent]` — 74.02% savings

Top Feature Combination of Size 6

- `[Compound_Sentiment, LIWC_Death, LIWC_Religion, LIWC_Friends, LIWC_Assent, LIWC_Nonflu]` — 74.12% savings

Top Feature Combination of Size 7

- `[Compound_Sentiment, LIWC_Family, LIWC_Death, LIWC_Swear, LIWC_Friends, LIWC_Sexual, LIWC_Assent]` — 74.05% savings

Top Feature Combination of Size 8

- `[Compound_Sentiment, LIWC_Family, LIWC_Death, LIWC_Swear, LIWC_Religion, LIWC_Ingest, LIWC_Friends, LIWC_Assent]` — 73.95% savings

Top Feature Combination of Size 9

- `[Compound_Sentiment, LIWC_Family, LIWC_Death, LIWC_Swear, LIWC_Religion, LIWC_Ingest, LIWC_Friends, LIWC_Assent, LIWC_Nonflu]` — 73.95% savings

Top Feature Combination of Size 10

- `[Compound_Sentiment, LIWC_Family, LIWC_Death, LIWC_Swear, LIWC_Religion, LIWC_Ingest, LIWC_Friends, LIWC_Sexual, LIWC_Assent, LIWC_Nonflu]` — 73.83% savings

These results suggest that combining high-performing individual features enables even more targeted and efficient search, while retaining interpretability due to their linguistic grounding.

Analysis

This approach performs especially well when the source and target subreddits share strong psychological or topical similarities. Features like compound sentiment, references to religion, death, family, and social behavior frequently aligned with meaningful community overlap, allowing the A* search to efficiently prune the search space.

The single-feature heuristics already led to large reductions in node expansions, with the best individual feature (compound sentiment) reducing node visits by over 71% compared to BFS. However, combining features yielded even stronger results, with select combinations exceeding 74% savings. These gains demonstrate that small sets of linguistically meaningful dimensions can serve as effective and explainable proxies for subreddit similarity.

Conclusion

This project shows that semantic heuristics based on subreddit post features can dramatically improve pathfinding efficiency in large hyperlink networks. By using A* search guided by cosine distance over individual post-level features, I consistently reduced node expansions compared to BFS, with several features, such as Compound VADER Sentiment, LIWC_Family, and LIWC_Death, achieving over 71% average savings.

Combining features led to even better results. Certain groups of 4 to 7 features pushed savings past 74%, outperforming any individual dimension. These results highlight the value of interpretable, content-grounded heuristics for guiding search in complex networks.

On the technical side, I learned how to use feature-driven reasoning with classical search methods, run large-scale randomized experiments, and measure performance across thousands of real subreddit pairs.

Overall, this approach provides a lightweight and interpretable method for navigating subreddit networks. It sets the stage for possible work on dynamic heuristics, personalized search strategies, and more sophisticated use of semantic features across social graphs.

References

[1] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Community Interaction and Conflict on the Web. *Proceedings of the Web Conference (WWW)*, 2018.