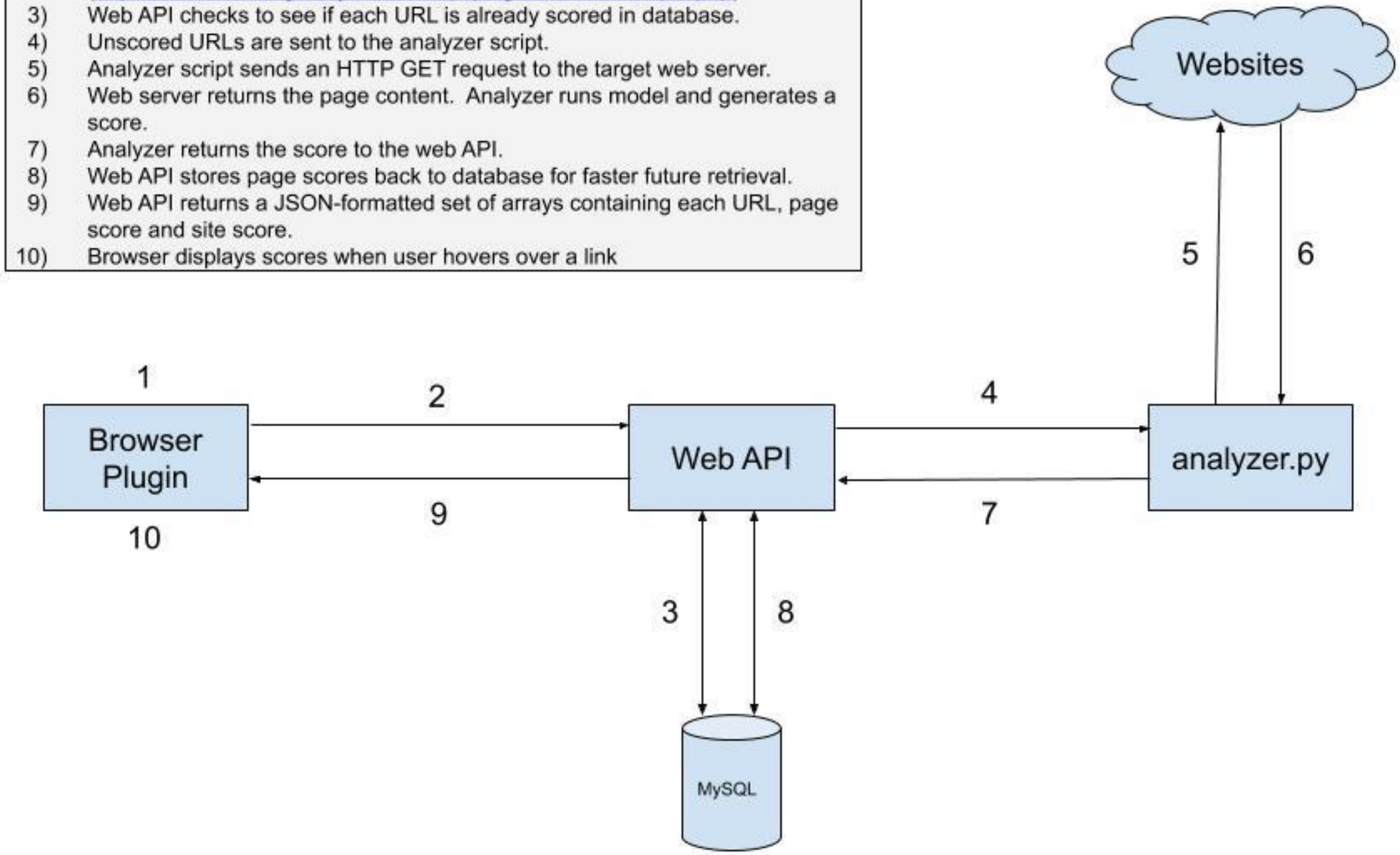
Web Crawling and Community Review to Prevent Misleading Links

Data Dogs - End of term status presentation

Data Dogs Analytics System Architecture

1) Browser plugin enumerates all links on the page.
2) Browser plugin submits JSON-formatted list of URLS to the web API. (https://www.datadoganalytics.com/api/plugin-submit-multiple.php)
3) Web API checks to see if each URL is already scored in database.
4) Unscored URLs are sent to the analyzer script.
5) Analyzer script sends an HTTP GET request to the target web server.
6) Web server returns the page content. Analyzer runs model and generates a score.
7) Analyzer returns the score to the web API.
8) Web API stores page scores back to database for faster future retrieval.
9) Web API returns a JSON-formatted set of arrays containing each URL, page score and site score.
10) Browser displays scores when user hovers over a link



API Features

Provides interface between browser extension, database, and analyzer

Returns page title, page score, and site score to browser extension

Caches recently-scored pages in database to reduce delay

Future API Features

Improved performance/caching to reduce user delay

Scheduled background job to periodically re-score indexed pages

Browser Extension Features

Sends URL of hovered link to web API for scoring

Appends response to page

Future Browser Extension Features

- Sidebar for comment viewing
- Content Extraction
- Link Color Highlighting
- (Possibly) Content Dashboard

Analyzer Features

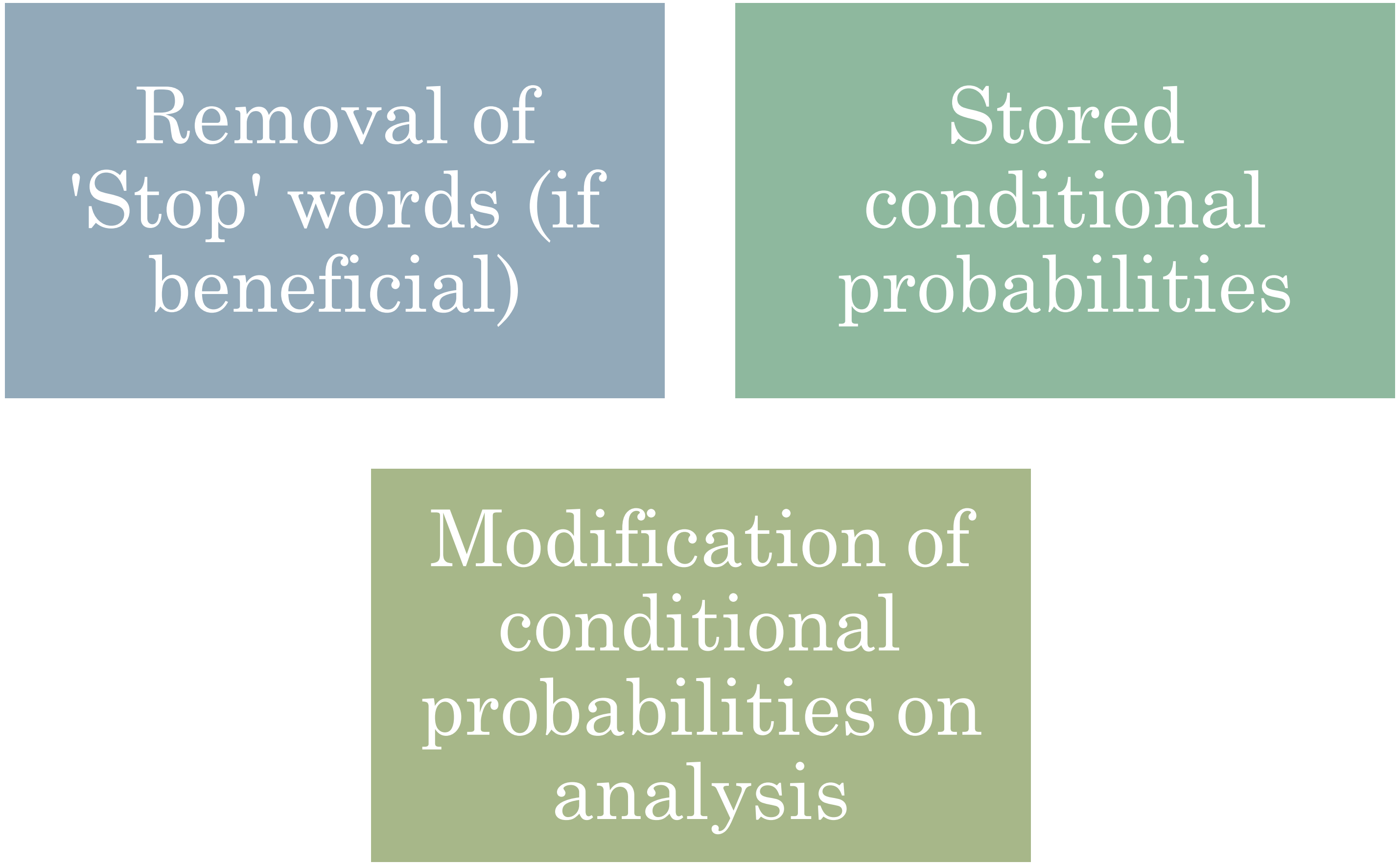
- Extracts title from URL
- Performs MNB classification on title
- Passes score and title to web API

The Algorithm

Uses probabilistic classification to determine headline tone

Conditional probabilities are obtained using frequency of word in document class.

Future Analyzer Features



Removal of 'Stop' words (if beneficial)

Stored conditional probabilities

Modification of conditional probabilities on analysis

Web Scraper Features

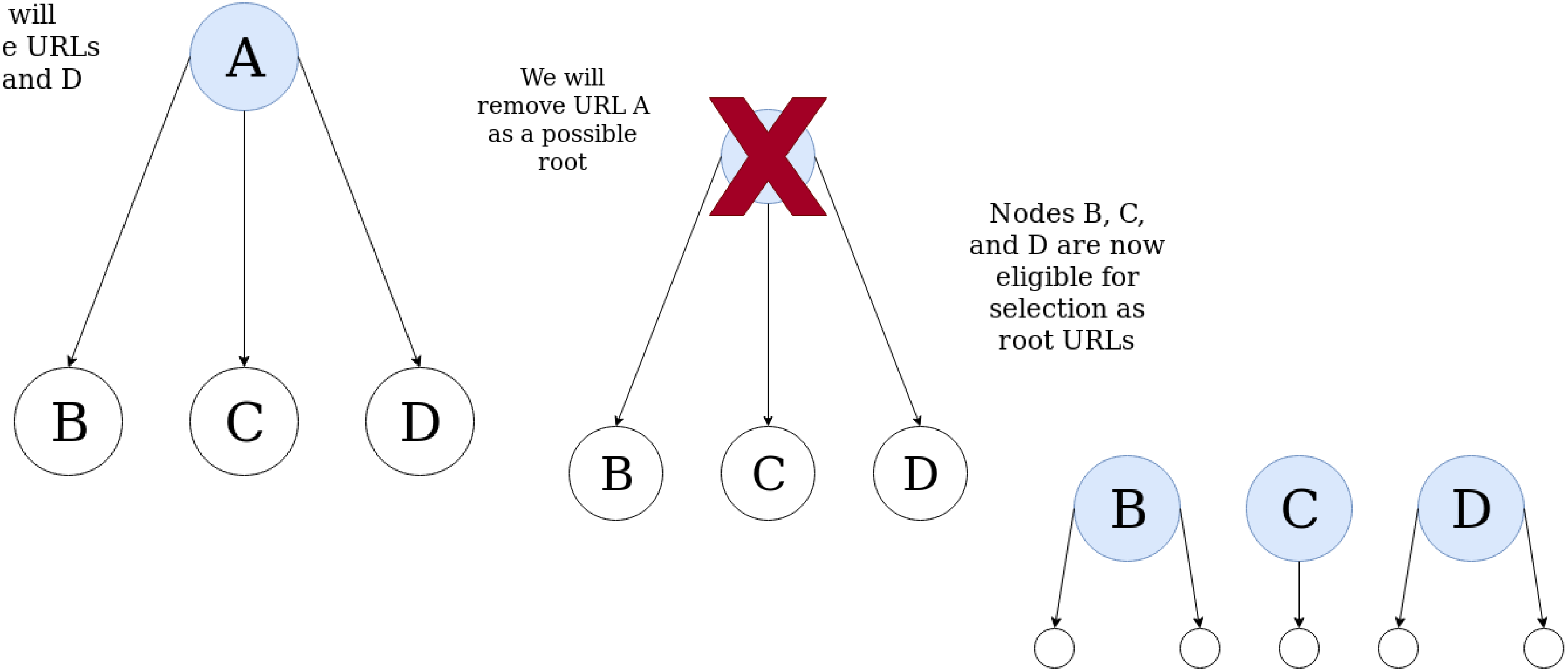
Currently a manual utility on the community forum.

Currently 'inconsiderate' (no throttling, does not process robots.txt)

Calls web API and provides articles for the user to view

Pages can be analyzed multiple times but cannot have their links analyzed multiple times.

Scraper Process



URL A is our current root. We will analyze URLs B, C, and D

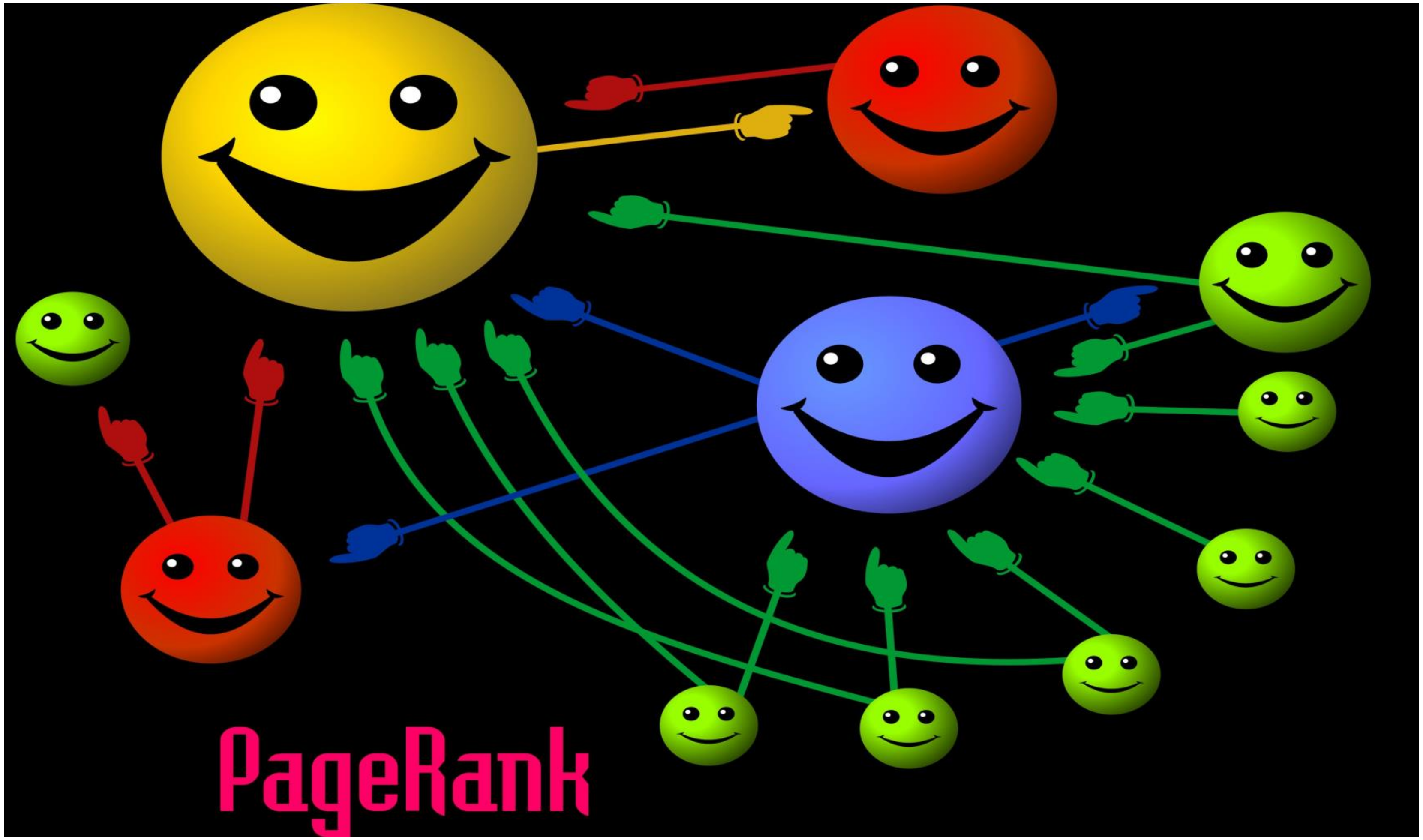
We will remove URL A as a possible root

Nodes B, C, and D are now eligible for selection as root URLs

Future Web Scraper Features

- Automation
- Strong consideration for other sites
- Client-side duplication prevention
- Live scraper updates (possibly)

Page rank



Page Rank Features

Takes URL as a command line parameter.

Searches through links that are linked to the parsed URL.

Returns a page score based on number of links that care linked to URL

Future Page Rank Features

IMPROVE RANKS ALGORITHM FOR SCORE PRECISION

IMPROVE URL LOOKUP

Community Forum Features

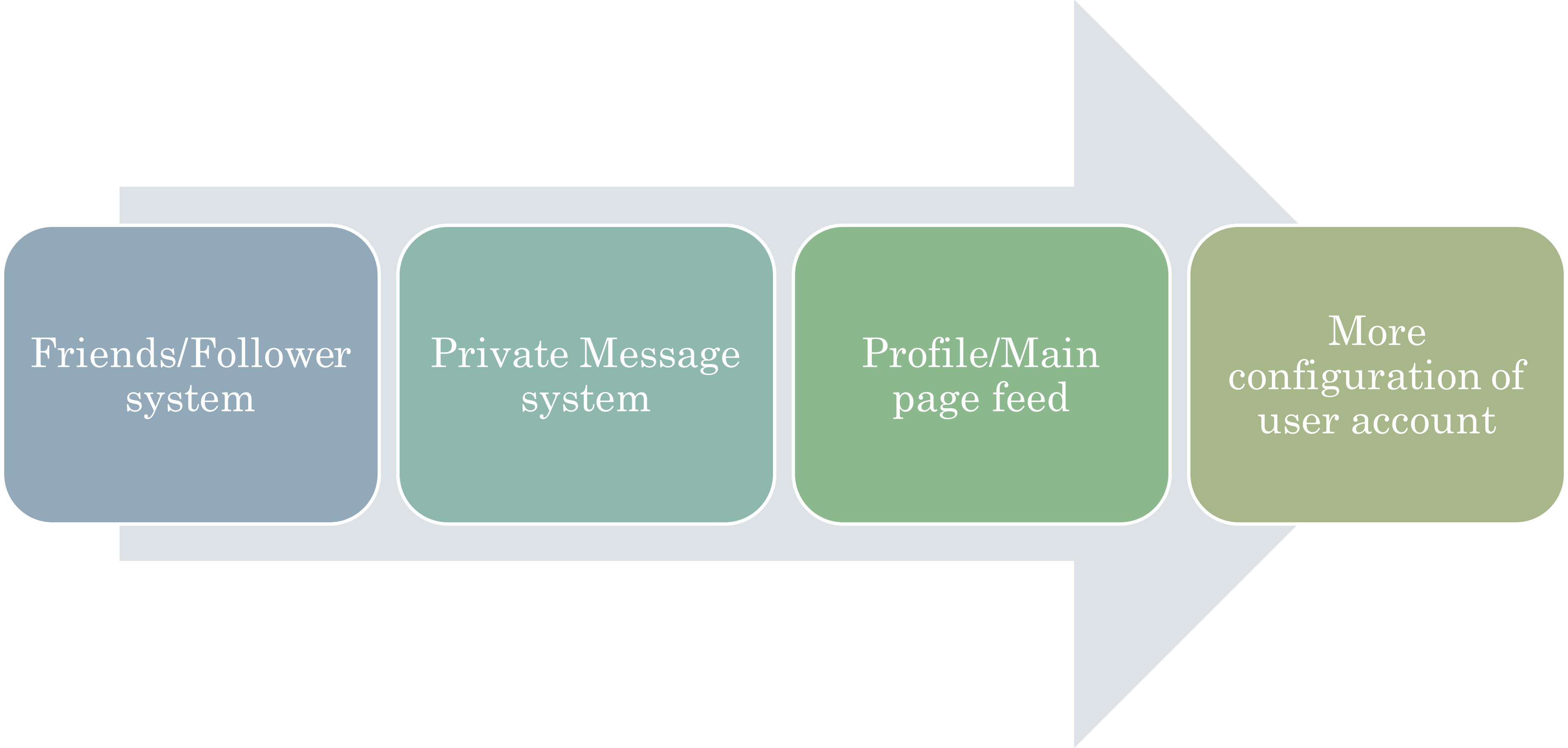
Users can search for analyzed articles in real-time

Users can manually scrape a URL to generate new articles.

Website Social Features

- Allows user account creation
- Own user profile page with personal profile configuration

Future Website Social Features



Friends/Follower system

Private Message system

Profile/Main page feed

More configuration of user account

First-term Progress

- All components communicate with each other
- Browser plugin displays page title, page score, and site score
- Bayesian and page rank analyzers are functional
- Web crawler code started
- Website with user accounts and ability to search scored pages
- User profile page with configuration ability
- Mechanism to manually re-score all indexed pages after updating analyzer code

Future Goals

- Improved analyzer results
- User reviews
- Improved website and browser plugin appearance
- Added social aspect
- Improved web crawler