**A General Simhash-based News Aggregator System**

**Pengcheng Hu1,a,\*, Xiangdong You1,b and Third Author Name2,c**

1Institute of Problem Solving, XYZ University, My Street, MyTown, MyCountry

2Department of Computing, Main University, My Street, MyTown, MyCountry

a email, b email, c email

\*corresponding author

**Keywords:**The News Aggregator, News Service, Simhash

**Abstract.** News aggregator is a system which aggregates online newspapers, blogs, videos from different sources, and recommend news to different users according to their interests. The system will crawl ten thousands of news from Internet every day, and try to recommend certain or similar news for users in a database which contains millions of news. We proposed a general Simhash-based aggregator system to handle common problem like depulication, Similarity and big event detection.

1. **Introduction (Heading 1)**

An emerging trend in modern press industry is providing digit news to readers. Users can browse news on a website belongs to a certain news service provider, however, the website just show the same contents to all users regardless of their interests and browsing history. Meanwhile, users need to browse different websites if he subscribes several news services.

A news aggregator is a system which aggregates online newspapers from different sources and provide personalized recommendation to users. The system need to crawl ten thousands of news from different news sources and clean them into well-formatted text. Then the system also need to calculate similarity between any two news so that it can recommend similar news to users according to their browsing history. A news aggregator can also provide special reports which contains series of news of a continuous event from different news provider. At the same time, news aggregator is an automatic system of news recommendation, it should be capable of detecting big events like natural disasters or political scandals when no editors involved in.

Be advised that papers in a technically unsuitable form will be returned for retyping. After returned the manuscript must be appropriately modified.

1. **Manuscript Preparation (Heading 2)**

We strongly encourage authors to use this document for the preparation of the camera-ready. Please follow the instructions closely in order to make the volume look as uniform as possible.

Please remember that all the papers must be in English and without orthographic errors.

Do not add any text to the headers (do not set running heads) and footers, not even page numbers, because text will be added electronically.

For a best viewing experience the used font must be Times New Roman, on a Macintosh use the font named times, except on special occasions, such as program code.

* 1. **Page Setup (Sub-Heading 2.1)**

The paper size must be set to A4 (210x297 mm). The document margins must be the following:

* Top: 3.5 cm;
* Bottom: 2.5 cm;
* Left: 2 cm;
* Right: 2 cm.

It is advisable to keep all the given values.

Regarding the page layout, authors should set the Section Start to Continuous with the vertical alignment to the top and the following header and footer:

* Header: 1.5 cm;
* Footer: 1.5 cm.

Any text or material outside the aforementioned margins will not be printed.

* 1. **First Section (Sub-Heading 2.2)**

This section must be in one column.

* + 1. **Paper Title**

Use 20-point type for the title, aligned to the center, linespace exactly at 14-point with a bold and italic font style and initial letters capitalized. No formulas or special characters of any form or language are allowed in the title.

Words like “is”, “or”, “then”, etc. should not be capitalized unless they are the first word of the title.

* + 1. **Subtitle**

Use 16-point type for the subtitle, aligned to the center, linespace exactly at 14-point with bold and italic font style.

Words like “is”, “or”, “then”, etc. should not be capitalized unless they are the first word of the subtitle. The initial letters should be capitalized. No formulas or special characters of any form or language are allowed in the subtitle.

* + 1. **Authors**

Author(s) name(s) should be aligned to the center with linespace single. The text must be set to 12-point and the font style set to bold.

There should be a spacing before of 6-point.

* + 1. **Affiliations**

Affiliation(s) should appear aligned to the center including organisation, address and e-mail.

The linespace must be single with 12-point of font size and the font style set to italic.

* + 1. **Keywords**

Each paper must have at least one keyword. If more than one is specified, please use a comma as a separator. Keywords should appear justified, with a linespace single, spacing before of 18-point, spacing after of 18-point and font size of 12-point.

The sentence must end with a period.

* + 1. **Abstract**

Each paper must have an abstract. The abstract should appear justified, with a linespace single, spacing before of 18-point and after of 60-point, and font size of 12-point.

The sentence must end with a period.

* 1. **Second Section**

This section must be in one column.

The section text must be set to 12-point, justified and linespace single.

Section, subsection and sub subsection first paragraph should not have the first line indent, other paragraphs should have a first line indent of 0,5-centimeter.

* + 1. **Section Titles**

The heading of a section title must be 12-point bold, aligned to the left with a linespace single and an additional spacing of 10-point before and 10-point after. The initial letters should be capitalized.

Dot should be included after the section title number.

* + 1. **Subsection Titles**

The heading of a subsection title must be 12-point bold with initial letters capitalized, aligned to the left with a linespace single and an additional spacing of 10-point before and 10-point after.

Words like “is”, “or”, “then”, etc should not be capitalized unless it is the first word of the subsection title.

Dot should be included after the subsection title number.

* + 1. **Sub Subsection Titles**

The heading of a sub subsection title should be in 12-point bold with initial letters capitalized, aligned to the left with a linespace single and an additional spacing of 10-point before and 10-point after.

Words like “is”, “or”, “then”, etc should not be capitalized unless it is the first word of the sub subsection title.

Dot should be included after the sub subsection title number.

* + 1. **Tables**

Tables must appear inside the designated margins.

It is advisable the use of text boxes in this case.

Tables should be properly numbered, centered and should always have a caption positioned above it. Captions with one line should be centered and if it has more than one line should be set to justified. The font size to use is 9-point. No bold or italic font style should be used. Spacing before and after should be of 12-point and 6-point, respectively.

The final sentence of a caption must end with a period.

Table 1 This caption has one line so it is centered.

|  |  |
| --- | --- |
| Example column 1 | Example column 2 |
| Example text 1 | Example text 2 |

Please note that the word “Table” is spelled out.

* + 1. **Figures**

Please produce your figures electronically, and integrate them into your document.

Check that in line drawings, lines are not interrupted and have a constant width. Grids and details within the figures must be clearly readable and may not be written one on top of the other.

Figure resolution should be at least 300 dpi.

Figures must appear inside the designated margins.

It is advisable the use of text boxes in this case.

Please do not use indentation and set the figure layout to in line with text.

Figures should be properly numbered, centered and should always have a caption positioned under it. Captions should be centered. The font size to use is 12-point. No bold or italic font style should be used. Spacing before and after should be of 6-point and 12-point, respectively.

The final sentence of a caption must end with a period.



Figure 1 This caption has one line so it is centered.



Figure 2 This caption has more than one line so it has to be set to justify.

Please note that the word “Figure” is spelled out.

* + 1. **Equations**

Equations should be placed on a separate line, numbered and aligned to the right. An extra line space should be added above and below the equation.

The numbers accorded to equations must appear in consecutive order inside each section or within the contribution, with number enclosed in brackets and set on the right margin, starting with the number 1.

The use of a table with two columns is advisable.

Example:

a = b + c (1)

* + 1. **Program Code**

Program listing or program commands in text should be set in typewriter form such as Courier New.

Example of a Computer Program in Pascal:

Begin

Writeln('Hello World!!');

End.

The text must be aligned to the left with the linespace set to single and in 12-point type.

* + 1. **Reference Text and Citations**

All references should be numbered in square brackets in the text and listed in the References section in the order they appear in the text.

References should be set to 10-point, justified, with a single linespace and hanging indent of 0,5-centimenter.

1. **Conclusions (Heading 3)**

We hope you find the information in this template useful in the preparation of your submission.

**Acknowledgements**

If any, should be placed before the references section without numbering.

**References**

[1] Maganioti, A.E., Chrissanthi, H.D., Charalabos, P.C., Andreas, R.D., George, P.N. and Christos, C.N. (2010) Cointegration of Event-Related Potential (ERP) Signals in Experiments with Different Electromagnetic Field (EMF) Conditions. Health, 2, 400-406.

[2] Bootorabi, F., Haapasalo, J., Smith, E., Haapasalo, H. and Parkkila, S. (2011) Carbonic Anhydrase VII—A Potential Prognostic Marker in Gliomas. Health, 3, 6-12.

**Appendix**

If any, the appendix should appear directly after the references without numbering, and not on a new page.