1. Analisis minimum support

* Pada apriori, dibutuhkan itemset yang muncul lebih dari 1 dokumen.
* Minimal minimum support = 2/x dokumen

2. Analisis topik dokumen tidak berhubungan (kasus 2^F tidak mungkin terjadi)

3. Fasilitasi nilai minimum support dapat diproses

4. Pemilihan minimum EO

[paper] : [1, 2] = 0.6931472

[text] : [0, 2] = 0.64520013

[construct] : [0, 2] = 0.64520013

[languag, system] : [0, 1] = 0.64520013

[system] : [0, 1] = 0.64520013

[languag] : [0, 1] = 0.64520013

[text, construct] : [0, 2] = 0.64520013

* Kalo entropy yang paling kecil sama, pilih yang paling panjang. Misal [text, construct] dipilih daripada [text] atau [construct] mengingat [text] atau [construct] himpunan bagian dari [text, construct]
* Kalau ada lagi yang sama dan panjang sama, dipilih indeks terakhir yang paling panjang

5. Konjungsi belum

6. Mengembalikan stemmed word ke kata dasar? (untuk cluster description)

6. Pemilihan kalimat:

HASIL SUMMARY:

|  |  |  |
| --- | --- | --- |
| ClusterDescription | Sebelum Di Summary | Setelah Summary |
| **[languag, system]** | While complete understanding of arbitrary input text remains in the future, it is currently possible to construct natural language processing systems that provide a partial understanding of text with limited accuracy. | While complete understanding of arbitrary input text remains in the future, it is currently possible to construct natural language processing systems that provide a partial understanding of text with limited accuracy. |
| While a computer program that can provide complete understanding of arbitrary input text remains a distant dream, it is currently possible to construct natural language processing systems that provide a partial understanding of certain types of text with limited accuracy. | In this paper, we describe our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in in- formation from multiple language sources. |
| In this paper, we describe our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in in- formation from multiple language sources. |  |
| In this paper, we describe our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in information from multiple language sources. |  |
| ------------------------ | -------------------------------------------- | ------------------------------------- |
| **[text, construct]** | While complete understanding of arbitrary input text remains in the future, it is currently possible to construct natural language processing systems that provide a partial understanding of text with limited accuracy. | While complete understanding of arbitrary input text remains in the future, it is currently possible to construct natural language processing systems that provide a partial understanding of text with limited accuracy. |
|  | While a computer program that can provide complete understanding of arbitrary input text remains a distant dream, it is currently possible to construct natural language processing systems that provide a partial understanding of certain types of text with limited accuracy. | In this paper, we propose a clustering algo- rithm, CBC (Clustering By Committee), in which the centroid of a cluster is constructed by averaging the feature vectors of a subset of the cluster members. |
|  | We present a clustering algorithm called CBC (Cluster- ing By Committee) that automatically discovers concepts from text. |  |
|  | In this paper, we propose a clustering algo- rithm, CBC (Clustering By Committee), in which the centroid of a cluster is constructed by averaging the feature vectors of a subset of the cluster members. |  |
|  | We presented a clustering algorithm, CBC, for automatically discovering concepts from text. |  |

Revisi:

1. Filter Concept: verb + noun -> lihat rezan
2. Konjungsi dibuang
3. User bisa memilih verb, adjective, noun untuk dimasukkan atau enggak
4. Lambda di MR pakai rumus -> lihat albert dan dimas (UDAH)

Pemilihan lambda:

Analisis:

Parameter memiliki range [0,1]. Lebih deket ke 0, lebih efektif dalam eliminasi redundansi. Karena nilai ini bergantung pada kumpulan kalimat yang ada, nilainya harus disesuaikan bergantung hal tersebut.

Sebagai contoh, kalo kalimat2nya banyak membahas hal berbeda, dan sangat sedikit overlap, maka ini musti diset mendekati 1 untuk mengurangi pengaruh ruas kanan persamaan MMR (artinya kita lebih membutuhkan ruas kiri/ lebih penting milih kalimat yg relevan daripada milih2 supaya gak redundan). Sebaliknya, kalo koleksi kalimat banyak overlap, nilai musti lebih kecil untuk mementingkan ruas kanan (supaya gak banyak redundan)

ANALISIS ACCEPTED!

Dari Mori & Takuro (Information Gain meets Maximal Marginal Relevance):