NAME

csb – クラスタリング

SYNOPSIS

```
#include <geta/cs.h>

cs_elem *csb(syminfo *q, int qlen, int qbias, WAM *wam, int dir_q, int cs_type,
        int *ncs, int thresh, int cswmax)
```

DESCRIPTION

csb は長さ *qlen* の行(列)IDリスト *q* をクラスタリングしてソート(デンドログラムソート)する．よって，*q* は破壊的に書きかえられる．*q* の各要素(*syminfo* 型)では，*id*(行(列)のID)を指定しておけばよい．*weight* に検索結果のスコアなどを指定しておくと，デンドログラムソートに反映される．クラスタリングの詳細についてはチュートリアル($GETASRC/tutorial/libcs 下)を参照．

ncs* で分割数(クラスタ数)を指定する．ncs* は実際に分割できた個数に書き換わる．各分割クラスタの詳細情報(*cs_elem* 型の配列)へのポインタが **csb** の戻り値である．

cs_type でクラスタリングアルゴリズムを指定する．現在，以下のアルゴリズムが実装されている．

```
CS_HBC          Hierarchical Bayesian Clustering
CS_WARD         Ward法
CS_GAVE         Groupe Average法
CS_SLINK        Single Link法
CS_CLINK        Complete Link法
```

qbias でバイアス情報を指定する．不明なら0を指定しておけばよい．

wam と *dir_q* はそれぞれ現在開いているWAMの識別子と向きである．

thresh で分割クラスタの最小要素数を指定する．**csb** は分割過程において要素数が *thresh* より少ないクラスタは無視する．

cswmax でクラスタ内単語リスト(後述) **csw* の最大長(単語数)を指定する．

RETURN VALUES

csb の戻り値は *cs_elem* 型の配列へのポインタである．各々の要素がクラスタに相当する．よって長さは実際に分割できた個数 **ncs* である．この配列の領域は **csb** が自動的に確保する．

cs_elem 型の定義は以下のとおり．

```
struct cs_elem {
  syminfo *csw;  /* pointer to word list */
  syminfo *csa;  /* pointer to article list */
  int cswlen;    /* length of csw */
  int csalen;    /* length of csa */
};
```

q を文書リストと仮定すると，**csw* はクラスタ内に含まれる単語のリストであり，重み順にソートされている．*cswlen* は単語の個数であり，引数で指定した *cswmax* 以下の値となる．この領

域は自動的に確保される．

*csa はクラスタ内に含まれる文書のリストであり，*csalen は文書数である．この領域も自動的に確保される．

何らかの理由で分割不能の場合 NULL を返す．ただし，この場合でもクラスタリングとデンドログラムソートは成功している場合が多い．

SEE ALSO

syminfo(3), free_cs_elems(3)