Dokumentumkategorizálás távközlési alkalmazásban

BENKŐ BORBÁLA KATALIN, PÁNDI ZSOLT

{bbenko, pandi}@hit.bme.hu

Kulcsszavak: hibatűrő rendszerek, trouble ticket, dokumentumkategorizálás, szövegfeldolgozás

Cikkünk kiindulópontját egy konkrét probléma, egyes trouble ticketing rendszerek tökéletlensége adta. A vizsgált trouble ticketing redszerek – noha igen hasznosak lehetnének a hálózat üzemeltetőjének és a hibatűrő hálózatokkal kapcsolatos kutatások szempontjából –, nem használhatók elég jól, mert a ticketek kategóriamezője csak kevés információt tartalmaz. Célunk az automatikus kategória detekció a ticketben található szöveges leírás alapján. Ehhez a dokumentumkategorizálás széles eszköztárát vonultatjuk fel. A konkrét probléma megoldásán túl áttekintő képet adunk a dokumentumkategorizálás jelenlegi helyzetéről, a szokásos módszerekről, a rendelkezésre álló eszközökről is. Ez különösen hasznos lehet olyan olvasóinknak, akik szintén szövegfeldolgozási feladatokkal (például szöveges üzenetek, ügyfélszolgálat levelek stb) szembesülnek.

1. Bevezetés

A hibatűrő hálózatokkal kapcsolatos kutatások egyik alapvető problémája, hogy a valós rendszerek rendelkezésre állását és megbízhatóságát jellemző paraméterekhez igen nehéz, gyakran lehetetlen hozzáférni. A berendezéseket és kábeleket gyártó cégeknek nem áll érdekükben az ilyen, verseny szempontjából érzékeny adatok nyilvánosságra hozatala, a szolgáltatók pedig vagy nem rendelkeznek az erre vonatkozó megfelelő statisztikai adatokkal, vagy ugyancsak üzleti titokként kezelik ezt az információt.

Távközlő hálózatok üzemeltetése során úgynevezett *trouble ticket* rendszerekben rögzítik az előforduló hálózati hibákat, azok (valószínűsíthető) okát, a javításhoz szükséges időt, és további kapcsolódó információkat. Néhány nyilvános kutatási célú nemzeti hálózat az Interneten is hozzáférhetővé teszi hálózati hibabejelentő, más néven *trouble ticketing* rendszerét. Bár garantáltan teljes képet nem, valós rendszerek üzemeltetéséből származó értékes adatokat mindenképpen szolgáltathat az ilyen jellegű adatbázisok szisztematikus feldolgozása.

A hálózati adatok statisztikai vizsgálatának kérdése a hálózatüzemeltetők szempontjából is felvetődik. Lényeges különbség a nyilvános adatbázisok és a távközlési szolgáltatók éles adatbázisai között tulajdonképp nincs, csupán annyi, hogy az üzemeltetőnek érdeke az adatok pontossága és teljessége.

Sajnos a jelenleg alkalmazott, publikusan is hozzáférhető adatokat tartalmazó rendszerek [1,2] néhány hiányossága miatt a statisztikai elemzés az ember által írt folyószöveg értelmezését teszi szükségessé. Valószínűleg az üzemeltetők által alkalmazott rendszerek hiányosságai is hasonlóak ahhoz, amelyek az Interneten hozzáférhetőek.

Ennek a nagy kihívást jelentő feladatnak a megoldásához, azaz az ember alkotta szövegek elemzéséhez ad segítséget a dokumentumkategorizálás, amely nyelvészeti és matematikai alapokra épülő módszereket használ. Cikkünkben bemutatjuk a trouble ticket rendszerekben fennálló problémát; áttekintő képet adunk a dokumentumkategorizálás eszközeiről; végül javaslatot teszünk, hogy a rendelkezésre álló eszközökből hogyan érdemes rendszert építeni a trouble ticketek elemzéséhez.

2. Trouble ticket

Egy trouble ticket rendszer bejegyzései tipikusan az *1. ábrán* látható releváns információkat tartalmazzák.

A ticketek tartalmának értelmezéséhez célszerű vázlatosan áttekinteni egy ticket életét, amely lehetőséget ad arra is, hogy az adathalmazból levonható következtetéseket megalapozzuk. Egy ticket létrehozása vagy egy észlelt hibaeseménynek, vagy pedig egy jövőbeli, várhatóan a hálózatra hatással levő (például kockázatnövelő) eseményről kapott információnak köszönhető. A megnyitás időpontja értelmezéstől függően megegyezhet az adott esemény észlelésének vagy aktuálissá válásának idejével, de lehet a jövőbeli információ rögzítésének ideje is. A lezárás ideje a hibaesemény el-

1. ábra Trouble ticket tipikus tartalma

Trouble ticket

- Ticket azonosító
- Megnyitás időpontja
- Státusz (nyitott/zárt)
- Kategória / Rövid jellemzés
- Lezárás időpontja
- Részletes szöveges leírás
  (a tickethez kapcsolódó események leírása kötetlen formában)

hárítási időpontja, a bejegyzés aktualitásának elvesztése, vagy a nyitva felejtett ticketek ellenőrzésének időpontja lehet tipikusan.

Egy ticket létrehozásakor nyitott állapotba kerül, jelezve azt, hogy az üzemeltető részéről valamilyen reakciót igényel (beavatkozást vagy akár fokozott készültséget). A ticket lezárása idején a ticket zárt állapotba kerül. Ennek az attribútumnak a jelentősége, hogy az üzemeltető a rendszerben a nyitott ticketekre szűrve gyorsan tud információt szerezni az aktuális problémákról.

A ticket kategóriája általában a feldolgozást megkönnyítendő, előre definiált hibaosztályok egyikébe történő manuális besorolás eredménye, míg a rövid jellemzés egy lényegretörő, átfogó listában is megjeleníthető emlékeztető. E két mező funkciója nem fedi egymást teljesen, továbbá megemlítjük azt is, hogy a kategóriák megfelelő meghatározása kulcsfontosságú lehet a hálózatról generálandó statisztikák számára.

A tickethez kapcsolódó események leírása adja a legtöbb és a legpontosabb információt, ám sajnos csak az ember számára. Gépiesített kategorizálásnál igen komoly feladatot jelenthet a folyószöveg bejegyzések értelmezése, főleg, ha formai szabályoktól mentesek a bejegyzések, és a használt nyelv sem mindig ugyanaz.

Problémánk, hogy a ticketekhez kapcsolódó eseményekről a kategória mező tartalma alapján az [1] hálózat esetében csak `outage` vagy `scheduled maintenance` besorolás áll rendelkezésre, a [2] hálózatnál pedig a kategória mező helyett rövid szöveges jellemzés szerepel. Ahhoz, hogy ennél pontosabb kategória besorolást kapjunk, dokumentumkategorizálásra van szükség.

3. Dokumentumkategorizálás

A dokumentumkategorizálás az 1960-as évek óta kutatott, de még mindig aktuális tudományterület [3], mely alapvetően két feladattal foglalkozik:

• **Dokumentumklaszterezés:** a dokumentumhalmazt hasonlósági alapon klaszterekbe soroljuk. A klaszterek automatikusan – és a dokumentumoktól függően dinamikusan – alakulnak ki. Bemenő adat a klaszter sugara (vágási határ) vagy a kért klaszterek száma.

• **Dokumentumkategorizálás:** a dokumentumhalmaz elemeit előre definiált kategóriákhoz soroljuk. (Zavaró, de sajnos ennek a feladatnak a neve megegyezik a tématerület nevével.)

Ipari alkalmazásoknál tipikusan kategorizálási feladatról van szó, ám a jól működő kategorizálás alapfeltétele a kategóriák – nem csak jó, hanem – tökéletes kialakítása. Ezért gyakori, hogy a kategóriák kialakítása előtt egy reprezentatív részhalmazon klaszterezést végeznek, és az itt tapasztaltak alapján alakítják ki – esetleg hierarchikusan – a kategóriákat. Így elkerülhető, hogy túl általános vagy túl specifikus kategóriák alakuljanak ki, hisz nyilván nem hordozna túl sok információt, ha a dokumentumok 80%-a ugyanabba a kategóriába tartozna.

Jelenleg is aktívan kutatott terület a *szcenáró-hozzárendelés,* ahol a dokumentumot valamilyen előre definiált forgatókönyvhöz soroljuk (például kábelhiba lépett fel, majd ki lett javítva).

Mivel még a dokumentumok teljeskörű automatikus megértésétől (szemantika stb.) nagy távol állunk, a dokumentumfeldolgozók egyszerűsített modellt használnak. A két fő *dokumentummodell* a következő:

• **Szavak halmaza** (set of words) **modell.** A dokumentumot szavak halmazának tekinti, a szavak közti relációt nem veszi figyelembe. Egyszerű, gyors és meglepően hatékony [4].

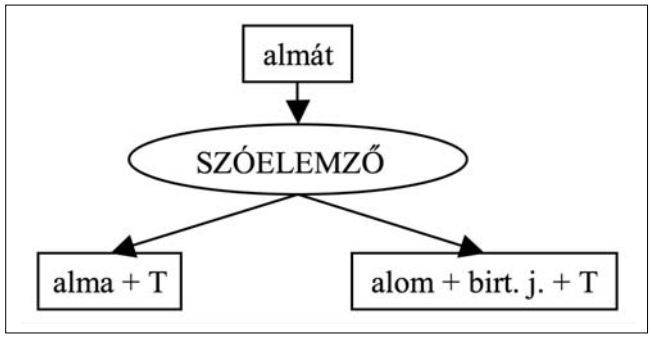
• **Nyelvészeti indíttatású** (linguistically motivated) **modellek.** Figyelembe veszi a szavak közti nyelvtani relációkat (alany-állítmány, tagadás, távoli kapcsolatok stb). Alapfeltétele egy jó nyelvtani elemző. Főleg szcenárió-hozzárendelésnél alkalmazzák, hisz itt nagy szükség van a viszonyok pontos azonosítására (megjavították vagy nem javították meg; ki okozta a kárt és ki hozta helyre) [5-7].

A gépi feldolgozás szempontjából alapvető különbség van a szeparáló és az agglutináló/flektáló nyelvek között. A szeparáló nyelvek (például angol) esetén a szóalak nem (illetve ritkán) változik, a nyelvtani reláció a szavak sorrendjében van kódolva. A dokumentum szavakra bontása közvetlen előállítja a szavak listáját. Az agglutináló vagy flektáló nyelveknél a felszíni szóalak és a szótő között akár nagyon nagy különbség is lehet. Ezért a dokumentum szólistájának előállításához szótövesítésre van szükség (morfológia). A szótövesítés sajnos újabb bizonytalanságot visz a rendszerbe, hisz a szótő igen gyakran nem egyértelmű *(2. ábra).* A modern szóelemzők közlik a szótő variánsok becsült valószínűségét is.

Egyes alkalmazásokban célszerű *hierarchikus kategorizálást* használni, azaz a dokumentumot előbb egy főkategóriába sorolni, majd azon belül finomítani. Ez nem csak a kategorizálás sebességét növeli meg (például 5 fő- és 10-10 alkategória esetén nem 50, hanem csak 15 illeszkedési vizsgálatot kell elvégezni), hanem a felhasználhatóságot is, hiszen így lehetőség lesz mind általános (főkategória alapján), mind specifikus (alkategória alapján való) szűrő kérdések megfogalmazására.

A kategorizálás szemantikai támogatása még gyerekcipőben jár, noha igen lényeges lenne. Néhány fontosabb kezdeményezés [8] és [9].

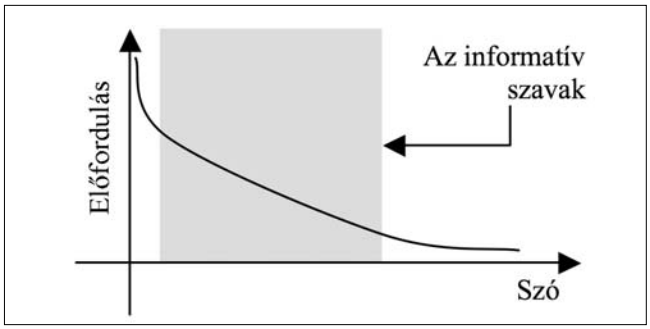
2. ábra
A szótövesítés bizonytalanságot visz a rendszerbe



3.1. „Szavak halmaza” modell

A „szavak halmaza” modell arra a feltételezésre épül, hogy a kategorizálási feladat jól megoldható csupán a dokumentumban előforduló (illetve a hiányzó) szavak alapján. A „szó” fogalmát szokás rugalmasan kezelni, például tulajdonnév vagy tagadás esetén (egy szónak számít: „Harry Potter”, „doesn’t repair”, „no problem”).

A *Luhn-megfigyelés*, vagy Luhn-szabály *(3. ábra)* szerint [10], ha a dokumentumban előforduló szavakat az előfordulások száma szerint sorbarendezzük, majd elhagyjuk a leggyakoribb és a legkevésbé gyakori szavakat, pont az informatív szóhalmaz marad meg. Heurisztikus magyarázat erre, hogy a nagyon gyakori szavak (névelők, kötőszavak) nem informatívak, a nagyon ritkák pedig valószínűleg nem kapcsolódnak szorosan a tárgyhoz.



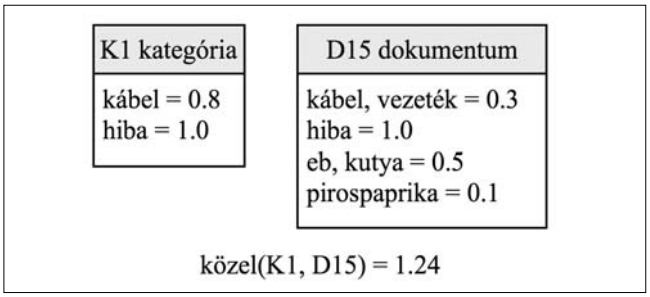
3. ábra A Luhn-megfigyelés

A Luhn-megfigyelés ebben a formájában nem használható a gyakorlatban (például egy 20 szavas dokumentum esetén nincs értelme gyakori és ritka szavakról beszélni). Ezért a szógyakoriságot egy nagyobb, reprezentatív korpuszon (szöveghalmazon) mérjük meg. Ez alapján készül el egy tiltólista (a vágandó szavak), illetve referencia szógyakoriságok. Fontos, hogy a mérési korpusz reprezentatív legyen (a feldolgozandó dokumentumokhoz hasonló elemekből álljon), hisz nyilván más mérési érték adódik általános szövegre, műszaki leírásra, dadaista versgyűjteményre stb.

A dokumentum és a kategória összerendelésére több különböző lehetőség van:

• **Bool-i összerendelés.** A dokumentumot akkor soroljuk a kategóriába, ha a kategória összes kulcsszavát tartalmazza. Ez több problémát is felvet, például mi történik, ha egy dokumentum az *x* és *y* kategória kulcsszavait egyaránt tartalmazza. Nem szokás használni.

• **Vektortér alapú összerendelés:** Egy vektort használ a dokumentum, illetve a kategória információtartalmának leírására. A vektor minden eleme egy szó súlyát jelenti (például az első elem a „kábelhiba” szóét). A kategória leíró vektorában a v_i elem a szó fontosságát jelenti; a dokumentum leíró vektorában pedig a szó jellemzőségét a dokumentumra (a szó előfordulási számán és a Luhn-szabályon alapuló normált érték). A dokumentum és a kategória közelségét egyszerű vektoriális szorzattal kapjuk meg (esetleg normálva). Lehetőség van tiltó szavak felvételére is (negatív súly a kategóriavektorban). A dokumentumot a hozzá legközelebb eső kategóriába soroljuk. A vektortér-modellbe könnyen beépíthető *rejtett szemantika* is, hisz csak annyit kell tenni, hogy a vektor i. eleme nem egy szót, hanem egy szinonima halmazt jelöl. A gyakorlatban természetesen a vektor helyett ésszerűbb adatstruktúrákat használunk.



4. ábra Vektortér alapú összerendelés

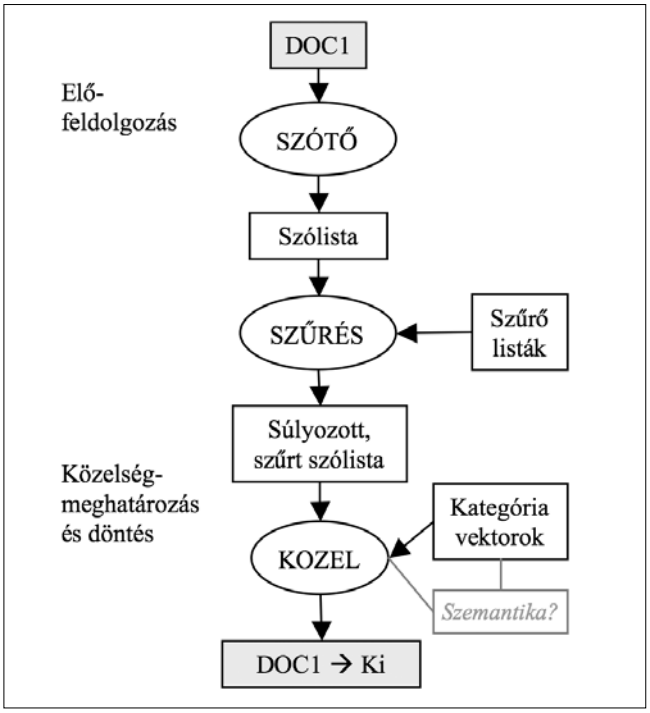
A *4. ábrán* egy normálás nélküli, rejtett szemantikus összerendelés látható. A közelség a mindkét helyen előforduló szavak súlyának összege.

Egy szó alapú kategorizáló tipikus architektúrája a *5. ábrán* látható. A dokumentumból először elő kell állítani a szavak listáját (általában szótövesítéssel), majd ezt a Luhn-szabály és/vagy szűrőlisták alapján redukálni és súlyozni. Ezután meghatározzuk a kategóriánkénti közelséget (lehetőleg valamilyen szemantikai támogatással), majd a dokumentumot a legközelebbi kategóriához rendeljük.

3.2. Nyelvészeti indíttatású modellek

A nyelvészeti indíttatású modellek alapötlete, hogy a kategorizáláshoz ne pusztán a szavakat, hanem a köztük levő relációt is vegyük figyelembe. A relációk azonosításához mondattani elemzésre van szükség, mely

5. ábra A szó alapú kategorizálás architektúrája



lehet mély elemzés (a teljes mondat elemzési fáját előállítjuk) vagy sekély elemzés (csak részlegesen elemzünk, például csak a névszói szerkezeteket).

A mondatelemzés alapján lehetőség van a mondat fő tartalmának beazonosítására. Például az esetek döntő többségében az alany, állítmány (és ha van, a tárgy) hordozza a fő információt. Ha elhagyjuk az egyéb mondatrészeket és a felesleges jelzőket és határozókat, pont a fő mondanivaló áll elő. Ezért gyakori, hogy a kategorizálás az alany-állítmány-tárgy hármas alapján történik, illetve kiegészítésként figyelembe veszik a hely és idő információt (ezek azonosítása sem mindig triviális, például „az előadás alatt" nem helyre, hanem időre utal).

Elvileg lehetséges lenne, hogy a kinyert információból egy háttérbeli szemantikai tudás *(domain knowlegde)* segítségével további következtetéseket vonjunk le, de ez ma még nem realitás (a probléma elsősorban nem a műszaki oldallal, hanem a háttértudás leírásával, illetve a következtetések értelmezésével van). A jelenlegi (kísérleti!) rendszereknél a leginkább szinonima relációt használnak.

Nemrég zárult le egy érdekes kutatási projekt [11], ahol gazdasági hírek feldolgozása volt a cél. A rendszer az adás-vétel eseményt képes felismerni akár eladásról, akár vételről szól a hír.

A nyelvészeti indíttatású dokumentummodellt komplexitása, mondattani elemző problémái, szemantikai háttérrendszer fejletlensége stb. miatt kategorizálásra olyan esetekben szokás használni, amikor a szó alapú modell nem hatékony. Ez elsősorban két esetet jelent:

– Homogén dokumentumhalmazt esetén. Itt a dokumentumok szókészlete kvázi-állandó (ugyanazok a szavak vannak az összes dokumentumban), ezért csupán a szavak alapján lehetetlen lenne kategorizálni.

– Ha a dokumentum szerkezete ezt indokolja (például sok tagadás).

4. Kategorizálás trouble ticketing rendszerekben

4.1. Milyen feladatot oldjunk meg?

Nézzük, milyen kategorizálási feladatnak van értelme trible ticketing rendszerekben.

• *Klaszterezés.* A dokumentumhalmaz előzetes analíziséhez tökéletes eszköz, de éles rendszerben értelmetlen lenne használni: nem rendel nevet a klaszterekhez, így a humán operátorok számára nehezen kezelhető; valamint új dokumentumok érkezésekor újra kéne klaszterezni, és lehetséges, hogy az új körben alapvetően más csoportok alakulnak ki.

• *Kategorizálás.* Egy jól megalkotott, (lehetőleg hierarchikus) kategóriarendszer nagyban megkönnyítené az operátorok munkáját. Egyaránt hasznos lenne az aktuális beavatkozások felügyeleténél (például a kategória alapján megállapítható a hiba súlyossága, ezáltal a beavatkozás sürgőssége), és a hálózati statisztikák készítésénél.

• *Szcenárióhozzárendelés.* A trouble ticket tartalma időben változik, nyilvánvalóan egy folyamatot ír le, azaz kézenfekvő folyamat-lefutási sablonhoz (szcenárióhoz) rendelni.

Bármely feladatot választjuk is, az biztos, hogy nem lehet teljes egészében az eredeti modellre hagyatkozni; figyelembe kell venni a trouble ticketek *nem időinvariáns* voltát (azaz hogy a ticketek tartalma és ezáltal a kategóriája az idő előrehaladtával változhat). A ticketeket minden új bejegyzés után újra kell kategorizálni.

4.2. Milyen modellt használjunk?

Tekintve, hogy a trouble ticketing rendszer alapvetően angol nyelven áll rendelkezésre, és az angol elsősorban szeparáló nyelv (ráadásul elég lokális természetű szerkezetekkel), elegendő a szóhalmaz modell. A tagadásokat és néhány egyéb fontos szófordulatot egy szónak tekintve elég jó modell adódik. A mondattani elemzés nagyon elbonyolítaná a rendszert, és minden valószínűség szerint kevesebb haszonnal járna, mint amennyi bizonytalanságot eredményez.

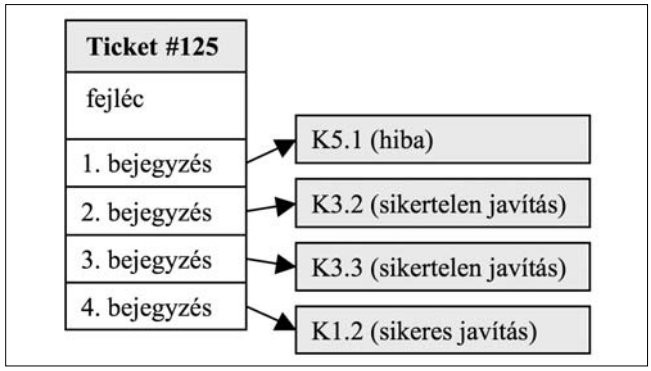
A másik kérdés, hogy pontosan mit akarunk kategorizálni, a trouble ticket bejegyzéseit, vagy az egész ticketet egyben. Az egyes bejegyzések kategorizálása mindenképp értelmesebb feladatnak tűnik. Később, egy újabb körben a bejegyzések kategóriái alapján már az egész ticketet is kategorizálhatjuk.

4.3. Rendszer vázlat

Első lépésként meg kell oldani a dokumentum (értsd: bejegyzések) szavakra bontását, majd elő kell állítani a szűrőlistákat, és a referencia szósúlyokat. A szómodell megalkotásához többféle segédeszközt is használhatunk (szűrők, named entity detector stb). Egy néhány ezer elemű dokumentumhalmaz alapján már nagy biztonsággal tudunk használható szűrőlistát készíteni és a referencia súlyokat számolni. Tekintve, hogy szakszövegről van szó, nem szerencsés a nyilvánosan hozzáférhető szósúlygyűjtemények használata.

A következő lépés a kategóriák megalkotása. Érdemes néhány próba klaszterezést futtatni (különböző vágási határokkal), majd az előálló klasztereket al-klaszterekre bontani. A klasztereket és alklasztereket áttanulmányozva képet kaphatunk a dokumentumok tartalmi eloszlásáról (vagy rossz esetben a szómodellünk, szű-

6. ábra Egy ticket bejegyzéseinek kategorizálása



rőlistáink hibáiról). A kategória leírások első változatát akár közvetlen is kinyerhetjük a klaszterekből (néhány egyszerű vektorművelet). A kinyert leírást emberek számára értelmezhető címkékkel kell ellátni, illetve szükség esetén tovább finomítani.

Ha a klaszterek nem az általunk kívánt elven alakulnak ki, bevethetünk néhány trükköt. Közismert, hogy a klaszterezés érzékeny az első néhány elemre; ezt a dokumentumok permutációjával kivédhetjük. A másik eszköz, hogy kiválasztunk néhány tipikus dokumentumot – minden tervezett kategóriából egyet-egyet – majd a kiválasztott dokumentumot megsokszorozzuk (például 800 példányban lemásoljuk). Ezzel tudjuk „kényszeríteni" a klaszterezőt, hogy a többi dokumentumot – ha lehetséges – ezen 800 elemű gócpontok köré rendezze.

Mivel már minden előfeltétel adott (szűrőlista, kategórialeírások), megkezdhetjük a dokumentumok (ticket bejegyzések) kategorizálását. A *6. ábrán* egy ticket bejegyzéseinek kategorizálása látható. Példánkban a kategóriák 2 szintű hierarchiát alkotnak (például a K3 a sikertelen javítást jelenti, a K3.2 és a K3.3 pedig ennek két külön alesete).

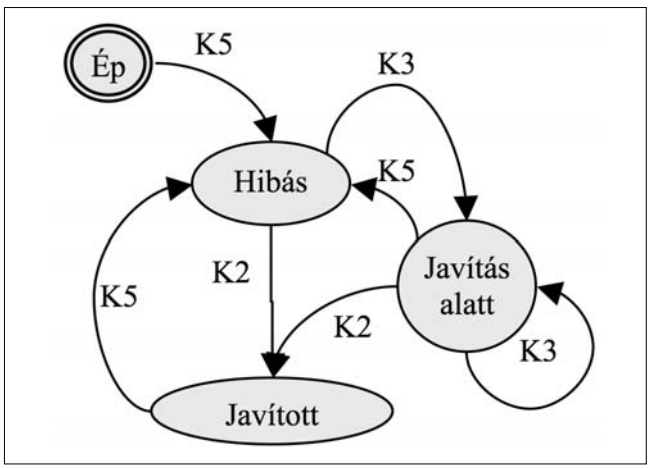
A kategorizált bejegyzések alapján előáll a ticket életút-leírása. Erre alapozva akár az egész tickethez rendelhetünk egy szcenárió mintát. Itt azonban az egyes bejegyzések sorrendje is lényeges.

Gyors és egyszerű egy véges automatát megalkotni, mely a bejegyzések kategóriája alapján rendel szcenáriót a tickethez. Például a *7. ábrán* látható véges automata részlet a hibabejelentések és sikeres/sikertelen javítások alapján vált állapotot. A bolyongás végállapota alapján kapjuk meg a szcenáriót. Az ábrán látható-nál bonyolultabb automatával lehetséges lenne olyan szcenáriók felismerése is, mint „többszörös javítási kísérlet után sem működőképes", „többszörösen javított, most működőképes" stb.

A véges automata modell különösen szerencsés választás, mert hosszú (sok bejegyzésből álló) ticketek esetén sem bonyolódik el.

Ha a bejegyzéseket meta-információval látjuk el (pl. idő, státusz), az operátor kényelmesen formálhat lekérdezéseket akár bejegyzés, akár ticket szinten. Sőt, mivel formalizált rendszert hoztunk létre, lehetőség van az adatok automatizált feldolgozására (például automatikus heti statisztika, általagos hibajavítási idő számítása a kábelhibákra stb).

7. ábra Véges automata a bejegyzések kategorizálására



5. Összefoglalás

Cikkünkben áttekintettük, hogy a modern dokumentumkategorizálási módszerek hogy használhatók fel egy aktuális telekommunikációs probléma megoldására.

Amellett, hogy a konkrét problémára vázoltuk a megoldás útját, cikkünk átfogó képet igyekezett adni a jelenleg rendelkezésre álló dokumentumkategorizálási eszközökről és a dokumentumkategorizálás jelenlegi helyzetéről (state-of-the-art).

A dokumentumkategorizálás alkalmazásával a trouble ticketing rendszerek használhatósága, ezáltal értéke megnő. Reményeink szerint ebből minden érintett fél profitál, és idővel jobb hálózat rendelkezésre állás érhető el.

Irodalom

[1] http://dooka.canet4.net
[2] http://www.switch.ch
[3] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Auckland, 1983.
[4] G. G. Chowdhury, 'Searching and retrieval', Introduction to Modern Information Retrieval, Library Association Publishing, London, 1999., pp.158–178.
[5] J. M. Ponte, Language modeling approach to information retrieval, In Proceedings of SIGIR'98, pp.275–281.
[6] D. Hiemstra, A Linguistically Motivated Probabilistic Model of Information Retrieval, 1998. In C. Nicolaou and C. Stephanidis (eds.) Proc. of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, ECDL-2, pp.569–584.
[7] B.K. Benkő, T. Katona, Information retrieval in homogeneous document sets using syntactical parse information. In Proceedings of the 9th International Symposium for Social Communication, Santiago de Cuba, 2005. ISBN 959-7174-05-7
[8] WordNet – http://wordnet.princeton.edu/
[9] OWL – http://www.w3.org/2004/OWL/
[10] H. P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development 2., 1958, pp.159–165.
[11] NewsML project – http://www.morphologic.hu