Item-válasz-elmélet alapú adaptív tesztelés

Item Response Theory based adaptive testing

ANTAL Margit[1], ERŐS Levente[2]

Sapientia EMTE, Műszaki és humántudományok kar, Marosvásárhely

[1]adjunktus, manyi@ms.sapientia.ro

[2]informatika szakos hallgató III. év, ideges@gmail.com

Abstract

One of the fastest evolving field among teaching and learning research is students' performance evaluation. Computer based testing systems are increasingly adopted by universities. However, the implementation and maintenance of such a system and its underlying item bank is a challenge for an inexperienced tutor. Therefore, this paper discusses the advantages and disadvantages of Computer Adaptive Test (CAT) systems compared to Computer Based Test systems. Furthermore, a few item selection strategies are compared in order to overcome the item exposure drawback of such systems. The paper also presents our CAT system along its development steps.

Összefoglaló

A diákok teljesítményének mérése a tanítás és tanulás kutatásának egyik legerőteljesebben fejlődő területe. A számítógépes tesztelést, illetve ennek adaptív változatát egyre szélesebb körben alkalmazzák a tudás felmérésére. Ennek ellenére egy adaptív tesztrendszer megvalósítása és karbantartása kihívást jelent a tapasztalatlan oktatók számára. Dolgozatunkban összehasonlítjuk a hagyományos és az adaptív tesztrendszereket, kiemelt figyelmet szentelve a tesztitemek kitettség-vizsgálatának, amely az adaptív rendszerek egyik hátrányos tulajdonságának tekinthető. Végül pedig bemutatjuk a saját adaptív tesztrendszerünket is.

Kulcsszavak: Item-válasz-elmélet, web alapú tesztelés, adaptív tesztelés.

1. Bevezetés

A diákok teljesítményének mérése a tanítás és tanulás kutatásának egyik legerőteljesebben fejlődő területe. A web alapú oktatási rendszerekbe integrált tesztrendszerek széles körben terjednek, és egyre több egyetemen fejlesztenek hasonló rendszereket [2], [3], [8]. A számítógépes tesztelés számos előnnyel rendelkezik a hagyományos, papír és ceruza alapú tesztekkel szemben, mint például: a tesztkérdésekhez különböző multimédia csatolható, a teszt kiértékelése azonnali, vagy például a gyakorló rendszerekben útmutatásokat, kisegítő utalásokat nyújthatunk a rendszer használóinak.

Ebben a dolgozatban áttekintjük az *Item-válasz-elmélet* alapú adaptív tesztelést, kitérve az előnyök és hátrányok tárgyalására is. A dolgozat második felében a saját rendszerünket mutatjuk be, illetve az itemek kitettségét vizsgáló szimulációkat ismertetjük. A dolgozatot a következtetések levonásával zárjuk.

2. Item-válasz-elmélet

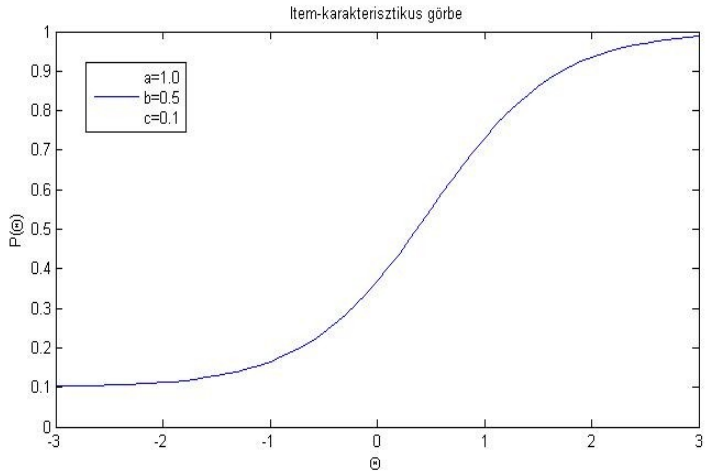
Az *Item-válasz-elmélet* egy valószínűségi tesztelmélet, amelynek fő célja a tesztitemek igazítása a vizsgázó képességi szintjéhez. A vizsgázó képességi szintjének becslése folyamatosan, a tesztelés során történik. Az adaptív tesztelés a következő lépésekből áll: (i) egy kezdeti képességszint beállítása (ii) az adott képességszinthez a legmegfelelőbb kérdés kiválasztása (iii) a képességszint újrabecslése a kérdésre adott válasz alapján. A második és a harmadik lépést addig ismételjük amíg a befejezési feltételek nem teljesülnek. Az adaptív tesztelés elméletének megalapozója Lord volt, aki 1952-ben arra a következtetésre jutott, hogy amíg egy személy képességi szintje teszt-független, addig a teszteredmények mindig teszt-függők [5]. A valószínűségi tesztelmélet fejlődésének következő

mérföldköve a Georg Rasch által 1960-ban ismertetett egyparaméteres logisztikai modell volt [10], amely a későbbiekben Rasch modellként vált ismertté. A következő évtizedeket a Item-válasz-elmélet alapú alkalmazások megjelenése jellemezte.

A következőkben a háromparaméteres logisztikai modellt mutatjuk be. Ebben a modellben minden egyes itemhez egy item-karakterisztikus görbét rendelünk, amely megmutatja, hogy adott Θ képességszintű diák milyen valószínűséggel válaszol helyesen az adott itemre. Az item-karakterisztikus görbe egyenlete a következő:

$$P(\theta) = c + \frac{(1 - c)}{1 + e^{-Da(\theta - b)}} \qquad (1)$$

ahol a az item diszkriminációja, b a nehézsége és c pedig a válasz kitalálásának valószínűsége. Az item nehézsége azonos skálán mozog a vizsgázó képességszintjével. Elméletileg ez a skála -∞ és +∞ között mozog, a gyakorlatban azonban elegendő -3 és +3 közötti intervallum [1]. A diszkrimináció azt mutatja meg, hogy az item mennyire jól választja szét a vizsgázókat az adott nehézségi szinten. Ez a paraméter a görbe meredekségét határozza meg annak középső szakaszában. Minél meredekebb a görbe ezen szakasza, annál nagyobb az item diszkriminációja az item nehézségi szintjén. A kitalálási faktor egy valószínűségi érték, például egy igen/nem választ váró kérés esetében értéke 0.5, a D pedig egy skálázási faktor, amelynek 1.7 értéket szokás használni. Az 1. ábra egy item-karakterisztikus görbét szemléltet.



1. ábra *Item-karakterisztikus görbe*

A logisztikai modell másik fontos eleme az item-információ függvény, amely méri, hogy az item segítségével mennyire lehet pontosan becsülni a képességszintet. Ha az item-információ nagy, nagyobb pontossággal lehet meghatározni egy item után a képességszintet, így a nagyobb információs mutatóval rendelkező kérdés kerül a vizsgázó elé, nem pedig az, amelyre nagyobb valószínűséggel tud válaszolni. Az item-információ számítására a következő képletet használtuk:

$$I_i = \frac{P_i'(\theta)^2}{P_i(\theta)(1 - P_i(\theta))} \qquad (2)$$

A $P_i'(\theta)$, a valószínűség elsőrendű deriváltja.

Az adaptív tesztelés harmadik lépése az új képességszint becslése az előzetesen megválaszolt itemek alapján. A szakirodalom erre több képletet is ajánl, amelyeket rendre kipróbálva, a Rudner [11] által ajánlott bizonyult a legmegfelelőbbnek:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} S_i(\hat{\theta}_s)}{\sum_{i=1}^{N} I_i(\hat{\theta}_s)} \tag{3}$$

ahol

$$S_i(\theta) = (u_i - P_i)\frac{P_i'}{P_i(1 - P_i)} \tag{4}$$

valamint u_i az i. kérdésre adott válasz alapján 1, amennyiben a válasz helyes, illetve 0 ellenkező esetben.

Az *Item-válasz-elmélet* két esetben mond csődöt, amikor minden kérdésre helyes, vagy minden kérdésre helytelen választ ad a vizsgázó. Ezeket a szélsőséges eseteket figyelni kell, és egy adott kérdésszám után le kell állítani a tesztelést. Bármilyen más eset esetében a megállási feltételt a standard hibához kötjük. A standard hiba a képességszint becslésének pontosságát jellemzi, ezért ha ez az érték egy küszöbérték alá csökken, leállíthatjuk a tesztelést [9].

A standard hiba kiszámításához felhasználjuk a teszt-információ függvényt, amelyet a következő képlettel számítunk:

$$TI(\theta) = \sum_{i=1}^{N} I_i(\theta) \tag{5}$$

Ezután a standard hibát pedig így számíthatjuk:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \tag{6}$$

A megállási feltételt a mi rendszerünkben a standard hibához, egy minimális, illetve egy maximális itemszámhoz kötöttük. A standard hibának 0.33, illetve ez alatti értéket szokás használni [11].

2.1. Előnyök és hátrányok

Az adaptív tesztelés legnagyobb előnye a megbízhatóság, illetve az a jellemzője, hogy képes igazodni a vizsgázó képességeihez, ennek következtében a nagyon jó képességű vizsgázókat nem untatja nagyon egyszerű kérdésekkel, a gyenge képességűeknek pedig nem tesz fel túl nehéz kérdéseket, amelyek semmilyen információt nem nyújtanának a vizsgázó képességét illetően. Ezen tulajdonság egyenes következménye, hogy a vizsgázó képességszintjét rövidebb idő alatt, kevesebb item segítségével képes megállapítani.

Az adaptív tesztelés előnyei mellett számos hátránnyal is kell számolni. Az első hátránynak az tekinthető, hogy a módszer nem alkalmazható abban az esetben ha minden kérdésre csak helyes, illetve csak helytelen választ ad a vizsgázó. Ezt a két esetet külön kell vizsgálni, és egy adott minimális kérdésszám után le kell állítani a tesztelést maximális, illetve minimális képességszinttel.

Egy másik hátránya ennek a tesztelési módszernek, hogy nem veszi figyelembe, hogy az itemek milyen témakörökhöz tartoznak. Bizonyos felmérések esetében viszont rendkívül fontos, hogy bizonyos témakörökből egységesen mérjen a teszt. Ennek a problémának a megoldására Huang [6] egy sajátos adaptív algoritmust javasolt. Egy másik megoldást Wainer és Kiley [14] javasolt, amelynek lényege itemcsoportok kialakítása fejezetenként. Ezen itemcsoportok egységként kezelendők, vagyis kiválasztás esetén a csoporthoz tartozó minden item felhasználódik a tesztelés során.

Az *Item-válasz-elmélet* egyik legkényesebb problémája az itemek kalibrációja, ami az itemek előzetes megfelelő mintacsoporton való kipróbálását jelenti. Ezután nyilván kiderülhet az itemről, hogy az nem megfelelő, nem differenciál kellőképpen. Miután kiszűrtük a nem megfelelő itemeket, következik az itembank ellenőrzése. Egy jó itembanknak témakörönként nehézség és diszkrimináció

szempontjából is megfelelő eloszlású itemeket kell tartalmaznia.

Ha már van egy megfelelő itembankunk, még mindig adódhatnak problémák a tesztelések során. Bizonyos itemek kitettsége túl magas lehet, illetve más itemek pedig mellőzve lehetnek a kiválasztások során. Ez annak tudható be, hogy egy adott képességszint esetében a következő item kiválasztása az item-információ függvény segítségével történik. Az item-információ függvényt kiszámítjuk az összes olyan itemre, amely még nem volt kiválasztva az adott tesztelés során. Ezen itemek közül pedig a legnagyobb item-információval bírót választjuk ki. Mivel különböző item-paraméter kombinációk azonos item-információ értékhez vezethetnek, feltevődik a kérdés, hogy ezek közül melyiket a legmegfelelőbb választatni. A [4], [12] dolgozatok item kitettséget szabályozó módszereket ismertetnek.

3. Adaptív tesztrendszer implementáció

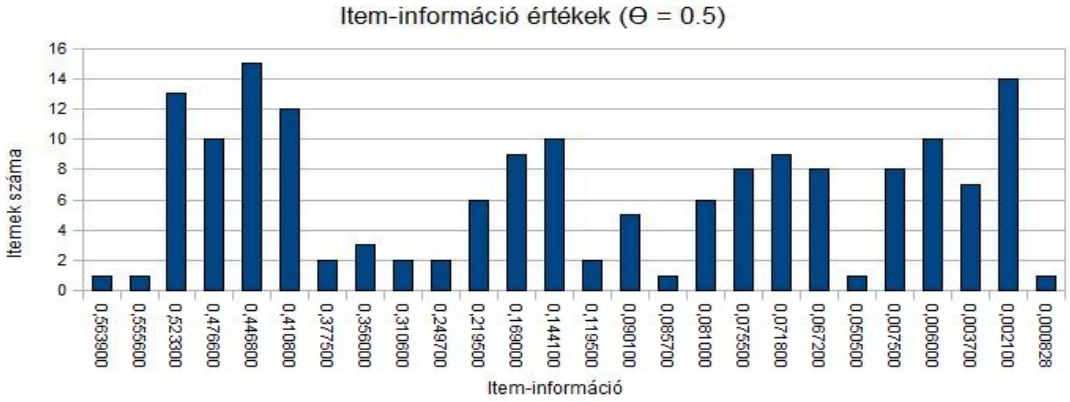
A következőkben a saját adaptív tesztrendszerünk fejlesztésének lépéseit ismertetjük.

3.1. Az item bank

Az itembankot 171 kérdés alkotja, amelyből 165 itemet saját fejlesztésű gyakorló tesztrendszerből vettünk át. Ebben a rendszerben az itemekhez öt féle nehézségi szint van rendelve, ezt skáláztuk a [-3,3] intervallumra. A kitalálási paramétert a helyes válaszok számának függvényében állítottuk be, például egy olyan kérdés esetében, ahol egy helyes választ öt lehetséges közül kell kiválasztani, a kitalálási paraméter értéke 0.2. A diszkriminációt nagyon nehéz becsülni, ezért ezt minden itemre egyforma 1-es értékre állítottuk. Ezt a paramétert csak megfelelő számú minta után lehet helyesen hangolni.

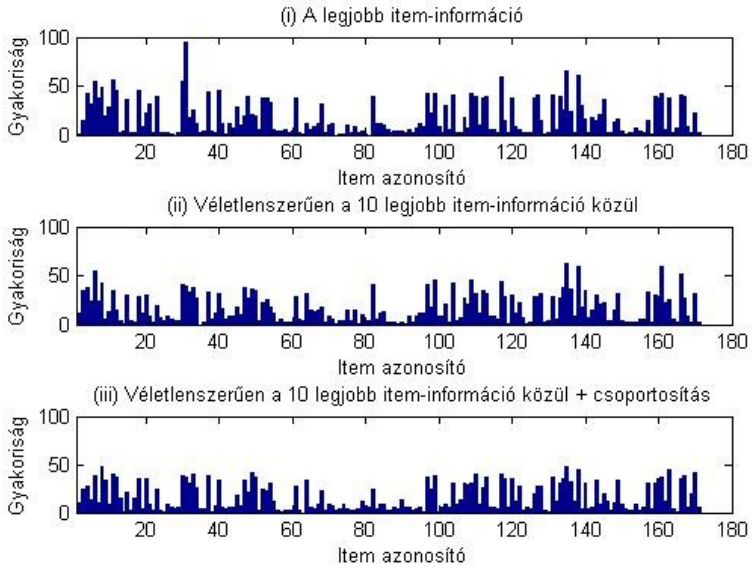
3.2. Itemek kitettségének szabályozása

Az implementációt megelőzően szimulációkat végeztünk, amelynek célja az itemek kitettségének szabályozása. A szimulációban 100 vizsgázót szimuláltunk, amely az előző fejezetben bemutatott item-információ függvény segítségével választotta ki a legmegfelelőbb kérdést az itembankból. A szimulációban három módszert vizsgáltunk: (i) a tesztelés során mindig a legnagyobb item-információt hordozó itemet választjuk (ii) a 10 legmagasabb item-információt hordozó itemből véletlenszerűen választjuk a következő itemet (iii) a hasonló item-információval rendelkező itemekből csoportokat alkottunk, majd ezen csoportokból véletlenszerűen választjuk ki a 10 legjobbat.



2. ábra *Item-információk klaszterek méretei*

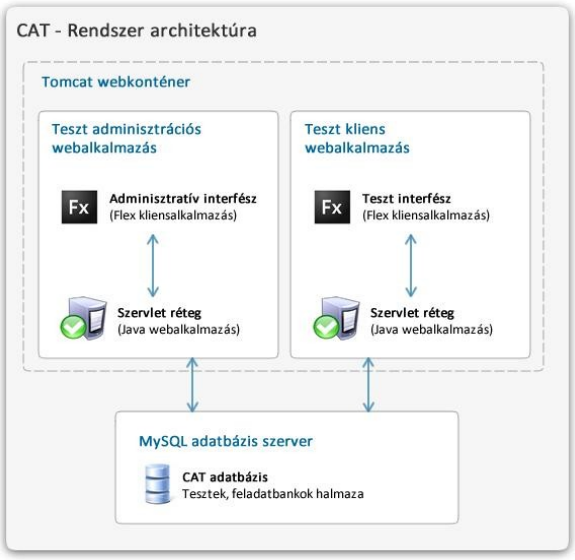
A 2. ábra a Θ=0.5 értékre keletkező csoportokat szemlélteti. Ebben az esetben a 10 legmegfelelőbbet úgy választjuk ki, hogy vesszük az első két csoportot (mindkettő egy-egy itemet tartalmaz), majd a harmadik csoportból, amelyet 13 item alkot kiválasztunk véletlenszerűen még 8 itemet. Az így kiválasztott 10 itemből ismét véletlenszerűen választunk egyet. A három módszerrel kapott item kitettségeket szemlélteti a 3. ábra, amelynek alapján a (iii) módszer szabályozza a legmegfelelőbben az itemek kitettségének mértékét.



3. ábra *Item-kitettség vizsgálata különböző szabályozó módszerekkel*

3.3. A rendszer architektúrája

Az adaptív tesztrendszerünket egy osztott rendszerként valósítottuk meg, amelyben szerveroldalon Java technológiákat használtunk, illetve Adobe Flex technológiát kliensoldalon. Az adatok tárolására MySql adatbázis-kezelőrendszert használtunk, amelyet a Hibernate perzisztencia keretrendszer állított elő az objektumorientált domain-modellből. A rendszerünk architektúráját a 4. ábra szemlélteti.



4. ábra *Adaptív tesztrendszer architektúra*

A teszt kliensalkalmazás a tesztek adaptív lebonyolításáért felelős, míg az adminisztrációs rész feladata az itembank karbantartása, a tesztek ütemezése, a teszt befejezési feltételeinek beállítása, illetve az elvégzett tesztekre vonatkozó statisztika készítése.

Az egyik leglényegesebb különbség a hagyományos és az adaptív tesztelés között az, hogy míg a hagyományos tesztelés esetében megengedhetjük a vizsgázónak, hogy visszalépjen az előzőleg

megválaszolt kérdésekhez, és módosítsa az előző válaszát, addig az adaptív változatban nincs visszalépés, hiszen minden kérdést az addig megválaszolt kérdések alapján megbecsült tudásszintnek megfelelően választottunk ki. Ha megengednénk a visszalépést, nagyon könnyen kijátszható lenne egy ilyen rendszer.

4. Következtetések

Dolgozatunkban bemutattuk az *Item-válasz-elmélet* alapú adaptív tesztrendszereket, majd részletesen tárgyaltuk a rendszer fejlesztése során felmerülő problémákat. Az itemek kitettségére először szimulációt végeztünk, majd a legmegfelelőbb módszert beépítettük egy osztott adaptív tesztrendszerbe. Annak ellenére, hogy jelen pillanatban még nincsenek méréseink a rendszer valós használatáról, a dolgozatban bemutatott elméleti rész és szimulációs eredmények jól hasznosíthatók egy ilyen típusú tesztrendszer implementálása során.

Terveink között szerepel egy item-kalibrációs modul elkészítése, amelyben a [7], illetve a [13] dolgozatokban bemutatott eredményeket szeretnénk megvalósítani.

Hivatkozások

[1] Baker, F. B. (2001). *The Basics of Item Response Theory.* ERIC Clearinghouse on Assessment and Evaluation. College Park, MD: University of Maryland.

[2] Barla, M., Bielikova, M., Ezzeddinne, A. B., Kramar, T., Simko, M., Vozar, O. (2010). On the Impact of Adaptive Test Question Selection for Learning Efficiency. *Computers & Educations 55*, 846-857.

[3] Baylari, A., Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications 36*(4), 8013-8021.

[4] Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment 5*(8).

[5] Hambleton, R. K., Jones, R. W., Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development, *ITEMS - Instructional Topics in Educational Measurement*.

[6] Huang, S. (1996). A Content-Balanced Adaptive Testing Algorithm for Computer-Based Training Systems. In Frasson, C., Gauthier, G., Lesgold, A. *Intelligent Tutoring Systems* (pp. 306-314). Third International Conference, ITS'96, Springer.

[7] Linden, W. J., Glas, C. A. W. (2006). *Capitalization on Item Calibration Error in Computer Adaptive Testing*, LSAC Research Report, 98-04.

[8] Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education 43*, 109-123.

[9] Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum.

[10] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

[11] Rudner, L. M. (1998). *An online, interactive, computer adaptive testing tutorial.* http://echo.edres.org:8080/scripts/cat/catdemo.htm

[12] Stocking, M.L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm. Technical Report RR 3-2. Princeton, New Jersey: Educational Testing Service.

[13] Stocking, M.L. (1990). Specifying optimum examinees for item parameter estimation in Item Response Theory. *Psychometrika 55*(3), 461-475.

[14] Wainer, H., Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement 24*, 189-205.